

Minería de Datos y Modelización Predictiva (I)

Fernández Hernández, Alberto. 54003003S

12/01/2021

Contents

1. Depuración de los datos	2
1.1 Introducción al objetivo del problema y las variables implicadas	2
1.2 Valores erróneos o no declarados	3
1.3 Análisis de valores atípicos	3
1.4 Análisis de valores missings (NA). Imputaciones	4
1.5 Relaciones con las variables input y objetivo	8
1.6 Transformaciones de variables y relaciones con las variables objetivo	9
2. Construcción del modelo de regresión lineal	10
2.1 Selección de variables clásica	11
2.2 Selección de variables aleatoria	12
2.3 Selección y justificación del modelo ganador	13
2.4 Interpretación de los coeficientes de dos variables	15
3. Construcción del modelo de regresión logística	15
3.1 Selección de variables clásica	16
3.2 Selección de variables aleatoria	18
3.3 Selección y justificación del modelo ganador	18
3.4 Selección del punto de corte óptimo	20
3.5 Interpretación de los coeficientes de dos variables	21

1. Depuración de los datos

1.1 Introducción al objetivo del problema y las variables implicadas

El objetivo principal del problema consiste en **obtener un modelo tanto de regresión lineal como de regresión logística que permita calcular no solo el porcentaje de votos a la derecha en un municipio (Dcha_Pct), sino además predecir si en un municipio habrá una mayoría o no de votos a la derecha (campo Derecha)**. Para ello, contamos con las siguientes columnas con la información de los municipios en España (8119 observaciones):

```
["Name", "CodigoProvincia", "CCAA", "Population", "TotalCensus", "Dcha_Pct", "Derecha", "Age_0-4_Ptge", "Age_under19_Ptge", "Age_19_65_pct", "Age_over65_pct", "WomanPopulationPtge", "ForeignersPtge", "SameComAutonPtge", "SameComAutonDiffProvPtge", "DifComAutonPtge", "UnemployLess25_Ptge", "Unemploy25_40_Ptge", "UnemployMore40_Ptge", "AgricultureUnemploymentPtge", "IndustryUnemploymentPtge", "ConstructionUnemploymentPtge", "ServicesUnemploymentPtge", "totalEmpresas", "Industria", "Construccion", "ComercTTEHosteleria", "Servicios", "ActividadPpal", "inmuebles", "Pob2010", "SUPERFICIE", "Densidad", "PobChange_pct", "PersonasInmueble", "Explotaciones"] => Total variables: 36
```

En primer lugar, y una vez eliminadas el resto de variables objetivo, debemos **recategorizar las variables cualitativas como factor**, dado que el formato establecido por defecto es numérico o cadena de caracteres. Debido a que contienen un número limitado de valores, existen un total de 4 variables categóricas, incluyendo la variable objetivo binaria (Derecha), así como los campos CodigoProvincia, CCAA, ActividadPpal y Densidad:

```
# c(2,3,7,29,33) -> (CodigoProvincia, CCAA, Derecha, ActividadPpal, densidad)
datos[,c(2,3,7,29,33)] <- lapply(datos[,c(2,3,7,29,33)], factor)
```

```
## Categorías CodigoProvincia: 52 ; CCAA: 19 ; Derecha: 2 ; ActividadPpal: 5 ; Densidad: 4
```

Por otro lado, podemos observar que los datos proporcionados **contienen un total de 34 variables independientes**, de las cuales cabe destacar el campo identificador *Name*, un campo con el nombre de cada municipio:

```
## Nombres de municipio unicos: 8102 de 8119 filas.
```

Salvo excepciones, en las que el nombre del municipio coincide, se trata de un campo que podríamos considerar como identificativo, por lo que no nos aportará información relevante al modelo y por ello lo eliminamos.

Por otro lado, nos encontramos con el campo *CodigoProvincia* que, a diferencia del anterior, el número de valores diferentes es significativamente menor (52 valores únicos). No obstante, nos encontramos ante la siguiente duda ¿Mantenemos el campo o lo eliminamos? Por un lado, recategorizarlo como una variable cualitativa puede llegar a entorpecer la elaboración del modelo, en especial si una o varias de las categorías no están lo suficientemente representadas y deben ser agrupadas. Además, nos encontramos con un segundo problema: **el código de provincia depende la CCAA asociada::**

```
sapply(datos[, c("CodigoProvincia")],function(x) Vcramer(x,datos$CCAA)) # Correlacion perfecta (1)
```

```
## CodigoProvincia
## 1
```

¿Cuál debemos eliminar? En primera instancia, deberíamos descartar aquella variable que esté menos relacionada con nuestras variables objetivo, según la V de Cramer:

```
sapply(datos[, c("CodigoProvincia", "CCAA")],function(x) Vcramer(x,varObjCont))
```

```
## CodigoProvincia      CCAA
## 0.5323588            0.5133780
```

```
sapply(datos[, c("CodigoProvincia", "CCAA")],function(x) Vcramer(x,varObjBin))
```

```
## CodigoProvincia      CCAA
## 0.7119410            0.7003718
```

Como primeros resultados, la V de Cramer obtenida nos indica que el código de la provincia está mejor relacionada con las variables objetivo. Sin embargo, debemos recordar el número de categorías de cada variable: en el caso de la Comunidad Autónoma, con 33 categorías menos la diferencia entre ambas relaciones es de apenas 0.011 en la variable objetivo binaria y de 0.018 en la variable objetivo continua, con muchas menos categorías, por lo que no parece ser tan necesario conocer de qué provincia proviene el municipio, sino que con la CCAA parece ser suficiente. Por otro lado, muchas de las categorías en CódigoProvincia podrían estar muy poco representadas. A modo de ejemplo, de las 52 provincias, 18 de ellas tienen menos de 100 valores en el conjunto de datos, lo que podría suponer no solo

recategorizarlas, sino además de ser un proceso computacionalmente más costoso de cara a la elaboración de los modelos (especialmente en la regresión logística):

```
sum(freq(datos$CodigoProvincia)$`n` < 100)
```

```
## [1] 18
```

Por otro lado, ¿Y si lo consideramos como variable numérica? ¿Mejora la V de Cramer?

```
## VarObjCont: 0.14539 ; VarObjBin: 0.2371896
```

Tampoco parece mejorar. Por tanto, dado que la diferencia V Cramer entre ambas variables no es tan significativa pese a aumentar el número de categorías, **elegimos la CCAA, por lo que eliminamos CodigoProvincia**. Antes de continuar, de cara a valorar la calidad de la depuración final guardamos en una variable los valores de correlación originales, con el objetivo de compararlos con los del conjunto de datos ya depurado:

```
corr.previa <- cor(datos[,unlist(lapply(datos, is.numeric))], use="complete.obs", method="pearson")
```

1.2 Valores erróneos o no declarados

A continuación, procedemos a eliminar aquellos valores no declarados en las variables, así como posibles valores fuera de rango:

1. *ForeignersPtge* negativos. Porcentajes de extranjeros menores a cero:

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -8.96    1.06    3.59    5.62   8.18   71.47
```

```
datos$ForeignersPtge<-replace(datos$ForeignersPtge, which(datos$ForeignersPtge < 0), NA) # Min < 0
```

2. Porcentajes de *SameComAutonPtge* y *PobChange_pct* superiores al 100 % (en el caso de *PobChange_pct* según la documentación son posibles los porcentajes negativos, aunque no menciona los porcentajes mayores a 100):

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   75.81   84.49   81.63   90.46   127.16
```

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

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -52.2700 -10.4000 -4.9600 -4.8974   0.0925  138.4600         7
```

```
datos$PobChange_pct <-replace(datos$PobChange_pct, which(datos$PobChange_pct > 100), NA) # Max > 100
```

3. Valores a 99.999 en la columna *Explotaciones*, posible indicativo de la ausencia de valores en estos casos:

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       22       52    2447    137    99999
```

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

4. Categoría “?” sin declarar en *Densidad*, por lo que lo recategorizamos a NA:

```
##      ? Alta Baja MuyBaja
## n    92.0 557.0 1053    6417
## %     1.1   6.9   13     79
## val%  1.1   6.9   13     79
```

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

```
## Recoded 92 values to NA.
```

1.3 Análisis de valores atípicos

Una vez corregidos los errores detectados, analicemos los valores atípicos más destacados empleando la función *describe*:

```
##              sd  skew kurtosis
## Population    46215.20 45.98 2814.43
## TotalCensus   34428.89 46.49 2888.34
## totalEmpresas  4219.37 53.68 3472.00
## ComercTTEHosteleria 1233.02 45.40 2646.94
## Servicios     2446.81 57.48 3830.74
## inmuebles     24314.71 44.53 2643.65
## Pob2010       47535.68 47.15 2939.56
## SUPERFICIE    9218.19  6.07  62.28
```

Como podemos observar en la salida anterior, las columnas con la población, el censo total, el número total de empresas, así como la superficie son los que mayor desviación presentan con respecto a su media, lo que se traduce, además de una elevada asimetría, **en indicios de la presencia de valores atípicos** (algo lógico si observamos las desviaciones típicas obtenidas, donde en algunas variables como en el caso de *Population* presentan desviaciones muy extremas, del orden de 46.000). Por ello, comenzamos analizando el porcentaje de valores atípicos en nuestro conjunto de datos (top 5):

```
##      Servicios      totalEmpresas      Population ComercTTEHosteleria
##      11.873383      10.506220      9.927331      9.853430
##      Pob2010
##      9.754896
```

En este caso, el máximo porcentaje corresponde con el campo *Servicios*, con un 11.87 %, además de que las variables con mayor porcentaje **corresponden con aquellas de elevada asimetría**. No obstante, ¿Es tan elevado el porcentaje de atípicos en cada columna? Veamos el porcentaje de *outliers* mediante la función *summary*:

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01232 0.65279 3.57739 8.86809 11.87338
```

Si nos fijamos en el tercer cuartil, podemos comprobar que no todas las columnas presentan un alto porcentaje de atípicos. De hecho, el 75 % de las columnas (26, aproximadamente) no supera el 10 % (menos de 800 filas en un conjunto de datos de más de 8.000), por lo que dada la proporción podemos considerar dichos valores como atípicos, recategorizándolos como *missing*:

```
## Total valores missing: 10064
```

1.4 Análisis de valores missings (NA). Imputaciones

Tras recodificar los valores atípicos como ausentes, debemos analizar la proporción de valores atípicos tanto por observación como por variable. Para ello, obtenemos el valor máximo de *missings* en ambos casos:

```
## Por.observation Por.variable
## 1              37.5          12.64
```

Aparentemente, mientras que el porcentaje de *missings* por variable es del 12.64 %, por observaciones detectamos un mayor número (37.5 %). No obstante, si empleamos la función *summary* en este último caso:

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.000 0.000 0.000 3.874 3.125 37.500
```

Vemos que el 75 % de las observaciones contienen aproximadamente un 3 % de valores *missings* o menos, por lo que no parece tratarse de varias filas (de hecho, solo el 25 % presenta un porcentaje de *missings* superior al 3 %, con una media muy pequeña en comparación con el valor máximo, lo que indica que se tratan de observaciones atípicas). Por otro lado, la pérdida de información en ambos casos no supera el 50 %, por lo que en lugar de eliminar las filas o columnas podemos imputarlos. No obstante, existen determinados campos que pueden ser imputados manualmente sin necesidad de emplear una media, mediana o de forma aleatoria:

1. *Age_19_65_pct*, cuyo porcentaje de edad puede calcularse a partir de la suma de *Age_under19_Ptge* y *Age_over65_pct* menos el 100 %:

```
x["Age_19_65_pct"] <- 100 - (as.numeric(x["Age_under19_Ptge"]) + as.numeric(x["Age_over65_pct"]))
# Ejemplo demostrativo de que 100 - Age_under19_Ptge + Age_over65_pct = Age_19_65_pct
```

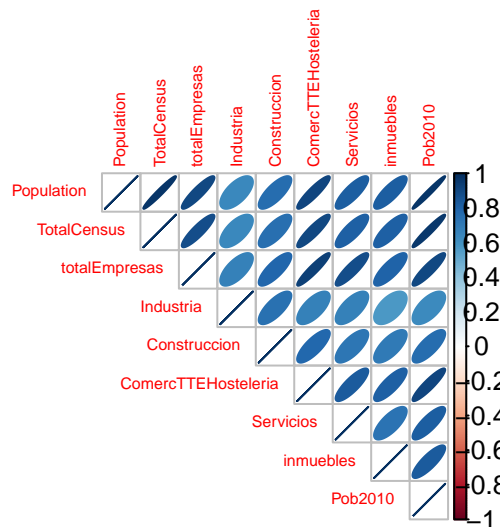
```
## Age_19_65_pct X100.Age_under_19.Age_over65
## 1      55.059      55.060
## 2      56.643      56.641
## 3      54.834      54.833
```

De hecho, solo se han encontrado 4 municipios en los que el $100 - (Age_under19_Ptge + Age_over65_pct)$ no coincide, por lo que parece tratarse de posibles casos aislados en los que los porcentajes no son complementarios:

```
sum(boxplot(abs((100 - (datos$Age_under19_Ptge + datos$Age_over65_pct)) - datos$Age_19_65_pct))$out > 1)
```

```
## [1] 4
```

2. *totalEmpresas* ¿Podría calcularse a partir de la suma del número de empresas de cada sector? *Industria, Construcción, ComercioTTEHostelería y Servicios*. Una primera prueba para comprobar si el campo *totalEmpresas* es la suma de cada columna podría ser analizando la matriz de correlación de valores *missings*: **Si falta cualquier sector, totalEmpresas tampoco debería aparecer al no poder calcularse:**



Efectivamente, **detectamos una correlación entre los valores *missing* de *totalEmpresas* con cada sector**. De hecho, no solo existe correlación entre *totalEmpresas*, sino incluso entre cada sector. A modo de ejemplo, si el número de empresas dedicadas al comercio no aparece, el número de empresas dedicadas a la construcción tampoco. Incluso si la población no aparece, tampoco suelen aparecer el número de empresas, el censo total e incluso tampoco la población registrada en el año 2010 (no se tratan de valores *missing* aleatorios, siguen un patrón de correlación). Por tanto, el campo *totalEmpresas* puede calcularse a partir de la suma de cada sector.

Como última prueba, realicemos una comprobación manual, sumando cada columna para comprobar si coincide con *totalEmpresas*:

```
x["totalEmpresas"] <- as.numeric(x["Industria"]) + as.numeric(x["Construccion"]) +
  as.numeric(x["ComercTTEHosteleria"]) + as.numeric(x["Servicios"])
# Ejemplo demostrativo de la suma de las empresas de cada sector, en funcion de ActividadPpal
```

```
##          ActividadPpal
## Coincide_Suma. ComercioTTEHosteleria Construccion Industria Otro Servicios
##          NO              0              0              0 3712              0
##          SI              1700             11              9 1215             195
```

Por lo general, cuando la actividad principal es *Otro*, la suma de cada columna no suele coincidir con *totalEmpresas* (al ser predominante otro tipo de Actividad, el resto de columnas valen 0). No obstante, de los valores *missing* de *totalEmpresas*, sólo existen 5 filas con la actividad principal a *Otro*, por lo que podemos realizar el cálculo manual sin afectar demasiado al conjunto de datos:

```
##
## ComercioTTEHosteleria      Construccion      Industria      Otro
##          527              0              0              5
```

```
## Servicios
## 326
```

Sin embargo, dado que existen valores *missing* tanto de *Industria*, *Construcción*, *Comercio y Hostelería* como *Servicios*, realizaremos el cálculo manual una vez imputados el resto de campos:

```
## Num. Missings Industria: 898 ; Construcción: 863 ; Comercio y Hostelería: 809 ; Servicios: 1026
```

3. *Densidad*, cuyo valor puede obtenerse a través del cociente entre *Population* y *SUPERFICIE*: si la proporción es menor a 1 decimos que la densidad es “MuyBaja”; si está entre 1 y 5 decimos que es “Baja” y si es mayor a 5 diremos que es “Alta”:

```
ifelse(proporcion < 1, densidad <- "MuyBaja", ifelse(proporcion >= 1 & proporcion <= 5,
  densidad <- "Baja", densidad <- "Alta"))
# Ejemplo demostrativo del cociente entre Population y SUPERFICIE a través de la función table
```

```
## Cociente_Pob_SUP
## Densidad Entre1Y5 Mayor5 Menor1
## Alta 0 140 0
## Baja 755 0 0
## MuyBaja 0 0 6152
```

De nuevo, dado que existen valores *missing* de *Population* y *SUPERFICIE*, realizaremos el cálculo de la Densidad una vez realizada la imputación en ambos campos:

```
## Num. Missings Population: 806 ; SUPERFICIE: 229
```

En relación con el resto de variables, debemos imputar sus valores. Dado que algunas de las variables presentan una elevada desviación típica, la media no parece ser la mejor alternativa, por lo que existen dos posibles opciones: mediante la mediana o de forma aleatoria junto con la mediana (dado que existen demasiados valores *missing* y no es posible con una sola imputación aleatoria). Para medir la calidad de ambas imputaciones, comparamos el porcentaje de correlación de los valores imputados en cada caso con los valores iniciales almacenados en *corr.previa*:

```
sum(abs(comparacion.corr.mediana) < 0.2) * 100 / (dim(comparacion.corr.mediana)[1] * dim(comparacion.corr.mediana)[2])
```

```
## [1] 80.97503
```

```
sum(abs(comparacion.corr.aleatorio) < 0.2) * 100 / (dim(comparacion.corr.aleatorio)[1] * dim(comparacion.corr.aleatorio)[2])
```

```
## [1] 81.92628
```

Podemos observar que aproximadamente **un 81.92 % de las correlaciones originales ha variado en menos de 0.2 con respecto a su correlación original al realizar una imputación aleatoria más mediana**, en comparación con un 80 % en el caso de la mediana. De esta forma, mediante la interpolación muchos de los valores *missings* quedan imputados. En relación a los valores faltantes, no tendría sentido realizar nuevamente una imputación aleatoria, ya que estaríamos asociando nuevos valores a partir de otro conjunto de puntos previamente imputado (*a grosso modo*, predecir a partir de una predicción). Por tanto, dada la menor variación que presenta en el porcentaje de correlación, **escogemos la imputación de forma aleatoria junto con la mediana**:

```
datos[,columnas] <- sapply(datos[, columnas],function(x) ImputacionCuant(x,"aleatorio"))
```

```
## 1218 valores missing tras la imputacion aleatoria (NO se han eliminado todos)
```

```
datos[,columnas] <- sapply(datos[, columnas],function(x) ImputacionCuant(x,"mediana"))
```

```
## 974 valores missing tras la imputacion con la mediana
```

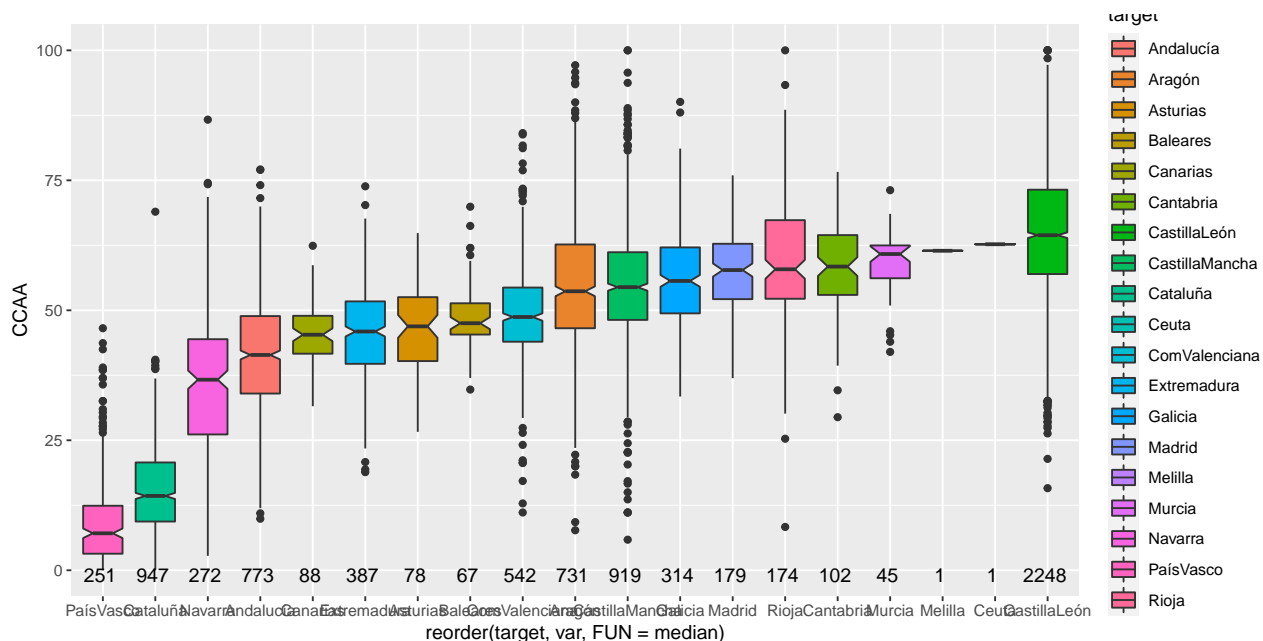
Como podemos observar, hemos conseguido reducir el porcentaje de *missings*. A continuación, si imputamos manualmente las tres columnas mencionadas anteriormente conseguimos reducir tanto el número de *missings* como el porcentaje máximo de atípicos:

```
## 0 valores missing. Columna con mayor % atipicos: 10.08745
```

Tras imputar las variables cuantitativas, debemos hacernos la siguiente pregunta. De las variables cualitativas ¿Podemos agrupar alguna de sus categorías? Salvo el campo *Densidad*, donde la frecuencia de cada categoría está repartida de forma equitativa:

```
##      MuyBaja Baja  Alta
## n      6509.0 1053 557.0
## %      80.2   13   6.9
## val%    80.2   13   6.9
```

Tanto en el campo *CCAA* como *ActividadPpal* debemos agrupar algunas de las categorías. Comenzando con las Comunidades Autónomas, disponemos de 19 valores diferentes, algunos de los cuales como Ceuta o Melilla con una única representación, tal y como se muestra a continuación (en la base de cada *boxplot* se encuentra el número de ocurrencias de cada categoría):



Para agrupar las Comunidades Autónomas, no solo agruparemos aquellas categorías con un menor número de variables sino además aquellas Comunidades **cuya amplitud en el diagrama de caja y bigotes sea similar**: País Vasco y Cataluña (PV_CAT); Navarra y Andalucía (AN_NA); ComValenciana, Extremadura, Asturias, Baleares y Canarias (CV_EX_AS_BA_CA); Aragón y Castilla la Mancha; (AR_CM) así como Galicia, Cantabria, Madrid, La Rioja, Ceuta, Melilla y Murcia (MA_CA_RI_CE_ME_MU_GA). De este modo, no sólo conseguiremos concentrar aquellas CCAA con una distribución de votos similar, sino además reducir el número de categorías. En el caso de Castilla y León, dado que se trata de la CCAA con mayor número de observaciones, no la agruparemos con otra categoría. No obstante, de cara a la creación de los modelos es importante tener en cuenta que se trata de la CCAA con la mayor distribución de votos hacia la derecha, además de ser la única categoría que no ha sido agrupada, por lo que lo consideraremos como la **categoría de referencia**, recodificando su nombre a *AA_CL* (de esta manera la categoría será elegida como referencia por orden alfabético):

```
##      AA_CL  AN_NA  AR_CM  CAT_PV  CV_EX_AS_BA_CA  MA_CA_RI_CE_ME_MU_GA
## n      2248.0 1045.0 1650.0 1198.0          1162.0          816.0
## %      27.7  12.9  20.3  14.8          14.3          10.1
## val%    27.7  12.9  20.3  14.8          14.3          10.1
```

En contraposición, nos encontramos con el campo *ActividadPpal*:

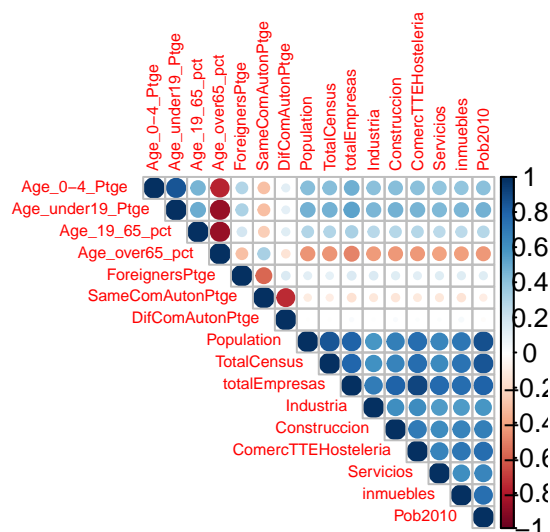
```
##      ComercTTEHosteleria Construcccion Industria  Otro Servicios
## n              2540.0          14.0          13.0 4932.0      620.0
## %              31.3           0.2           0.2  60.7       7.6
## val%           31.3           0.2           0.2  60.7       7.6
```

En este campo, las categorías *Construcccion* e *Industria* apenas tienen 14 y 13 apariciones, respectivamente. Por ello, dado que solo hay que agrupar dos categorías con poca representación, los agruparemos con la categoría principal: *Otro*, dado que la mediana en las tres categorías es muy similar, en relación al porcentaje de votos:

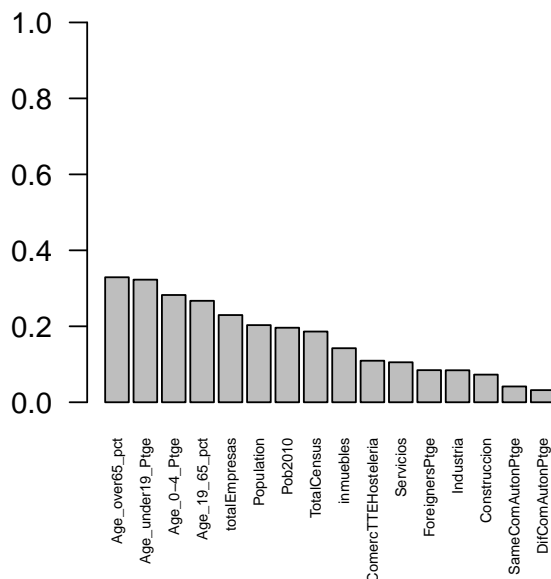
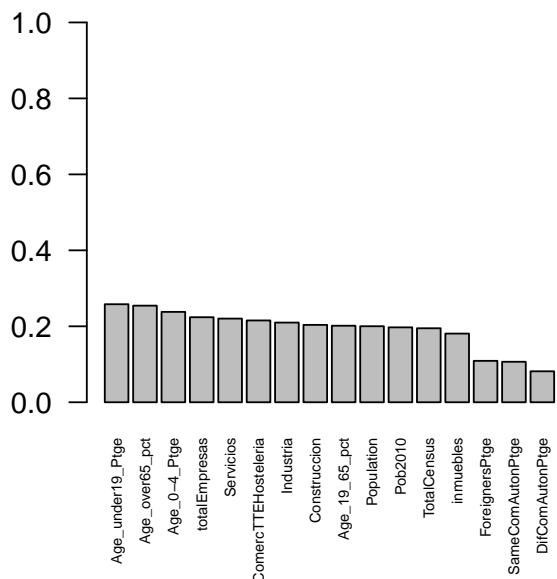
```
## Median (Otro): 56 ; Median (Construcccion): 56.004 ; Median (Industria): 57.353
```

1.5 Relaciones con las variables input y objetivo

Una vez recategorizadas las variables, ¿Cómo están relacionadas las variables *input* con las variables objetivo? O incluso algo más importante ¿Existe colinealidad entre las variables? Para responder a esta última pregunta, debemos analizar el siguiente subconjunto de la matriz de correlación:



Analizando el gráfico, debemos destacar tres grandes grupos de correlación. En primer lugar, **las edades**, donde cada porcentaje puede llegar a obtenerse (como pudimos observar con *Age_19_65_pct*) a partir del resto de edades, es decir, **son complementarios**. Por otro lado, **el porcentaje de personas que residen en la misma o diferente CCAA** (también complementarios), e incluso entre dichos porcentajes y el porcentaje de extranjeros. Como último bloque nos encontramos no solo con *totalEmpresas* y el resto de sectores, sino además con *Population*, *TotalCensus*, *inmuebles* y *Pob2010*, campos en los que pudimos detectar una elevada correlación entre valores *missing*. Por tanto, ¿Debemos eliminar algún campo? La respuesta es si, pero con cierto cuidado ya que solo podemos eliminar uno de cada grupo complementario (si eliminamos más de uno no podríamos volver a calcularlo). Para analizar qué campos podemos o no eliminar, realicemos la V de Cramer para ambas variables objetivo:



En ambos casos, de los campos de edad podemos eliminar *Age_19_65_pct* con menor significancia; del porcentaje de residencia en la misma o diferente CCAA podemos descartar *DifComAutonPtge*, lo que permitiría reducir el coste computacional para ambos modelos (a modo de variables *dummy*). ¿Qué podemos hacer con el bloque de *totalEmpresas*? Aunque debería eliminarse el menor número de variables posible, de cara a la regresión lineal y logística partiremos de un modelo general incluyendo todas las variables e interacciones, por lo que si mantenemos

todos los campos, el proceso de obtención será computacionalmente más costoso, por lo que se ha tomado la decisión de eliminar (del bloque *totalEmpresas*) la variable con el menor valor VCramer: *inmuebles* en el caso de la variable objetivo continua y *Construcción* en el caso de la variable objetivo binaria. El resto de campos los mantenemos.

Como primer análisis, el estudio realizado hasta el momento por el conjunto de datos, así como los V Cramer obtenidos, lleva a la conclusión de que tanto la CCAA como los porcentajes de edad (especialmente mayores de 65 y menores de 19 años), así como el número de empresas en el municipio parecen ser las variables más decisivas en relación al número de votos a la derecha.

1.6 Transformaciones de variables y relaciones con las variables objetivo

Tras eliminar las variables menos relevantes, debemos realizar las transformaciones de las variables continuas con el objetivo de que el modelo de predicción funcione mejor o pueda plasmar la verdadera relación con las variables objetivo:

```
input_cont<-data.frame(varObjCont,input_cont,Transf_Auto(Filter(is.numeric, input_cont),varObjCont))
input_bin<-data.frame(varObjBin,input_bin,Transf_Auto(Filter(is.numeric, input_bin),varObjBin))
```

La cuestión es ¿todas las transformaciones son significativas? ¿Aportan mejoría a los modelos? Computacionalmente sería muy costoso no solo trabajar con las variables originales, sino además con sus transformadas de cara a un modelo de regresión, además de que **las variables originales y sus transformadas están muy correlacionadas**. A modo de ejemplo, en *input_cont* la correlación mínima entre una variable original y su transformada es de 0.57 y en *input_bin* es prácticamente perfecta (1):

summary(vector.cont)						summary(vector.bin)					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.5708	0.7616	0.7989	0.8252	0.9578	1.0000	1	1	1	1	1	1

Por ello, comenzando con la variable objetivo cuantitativa filtraremos aquellas transformaciones cuya correlación con respecto a la variable objetivo mejore en más de 0.1 con respecto a la variable original, dado que (como podemos observar en el tercer cuartil del siguiente *summary*), sólo un 25 % de las variables originales ve mejorado su correlación en más de 0.1, por lo que en el resto de variables la mejoría es prácticamente nula:

```
correlaciones <- round(abs(cor(Filter(is.numeric, input_cont), use="pairwise", method="pearson"))[1,29:55])
summary(correlaciones)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01000 0.06000 0.06852 0.11000 0.20000
```

```
# Filtramos unicamente las transformadas que mejoren en mas de 0.1
input_cont <- input_cont[, !colnames(input_cont) %in% names(correlaciones[correlaciones < 0.1])]
input_cont <- input_cont[, c(-3,-4,-9,-12,-14,-15,-17,-19,-25)]
# Todas las transformadas significativas emplean escalas logaritmicas
names(correlaciones[correlaciones >= 0.1]) # Corresponden con las variables con mayor desv. tipica
```

```
## [1] "logxPopulation"          "logxTotalCensus"
## [3] "logxForeignersPtge"      "logxUnemployLess25_Ptge"
## [5] "logxUnemployMore40_Ptge" "logxAgricultureUnemploymentPtge"
## [7] "logxConstructionUnemploymentPtge" "logxtotalEmpresas"
## [9] "logxPob2010"
```

En relación con la variable objetivo binaria, para estudiar la importancia de las variables empleamos un criterio mucho más preciso que la V de Cramer: el criterio del **Valor de la Información**¹, una medida que permite analizar la influencia o poder predictivo que presenta una variable sobre otra dicotómica, por lo que cuanto mayor sea su valor de información o IV (generalmente a partir de 0.1), se dice que su poder predictivo es fuerte o influyente. En este caso, al igual que en la matriz de correlación restaremos los valores de información tanto de las variables transformadas como originales con el objetivo de analizar si la mejora es o no significativa:

```
salida.woe <- woebin(input_bin, "varObjBin", print_step = 0) # library scorecard
```

¹https://docs.tibco.com/pub/sfire-dsc/6.5.0/doc/html/TIB_sfire-dsc_user-guide/GUID-07A78308-525A-406F-8221-9281F4E9D7CF.html Se ha empleado dicho criterio ya que la V Cramer devolvía la misma importancia entre las variables originales y sus transformadas

```
summary(sapply(salida.woe[c(31:57)], function(x) x$total_iv[1]) -
        sapply(salida.woe[c(2:21,23:25,27:30)], function(x) x$total_iv[1]))
```

```
##      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
## -0.0167939 -0.0030499 -0.0008813 -0.0004858  0.0009320  0.0376600
```

Como podemos observar a partir de la salida anterior, un 75 % de las variables transformadas **ve mejorado su valor de información en 0.001 o menos**, e incluso empeora ligeramente en algunos casos. Por otro lado, el aporte máximo al valor de información ha sido algo más de 0.01, un valor muy poco significativo. Si a ello le añadimos que las transformadas son de tipo “x”, es decir, su valor multiplicado por 1.0001 (funcionesRosa.R), las descartamos del modelo de regresión logístico:

```
names(input_bin)[40:43] # Ejemplo del tipo de transformacion
```

```
## [1] "xSameComAutonDiffProvPtge" "xUnemployLess25_Ptge"
## [3] "xUnemploy25_40_Ptge"      "xUnemployMore40_Ptge"
```

Finalmente, una vez completado el proceso de depuración ya tenemos nuestros conjuntos de datos preparados para elaborar los modelos de regresión lineal y logísticos:

```
## Numero de columnas finales en input_cont: 31 ; input_bin: 31
```

2. Construcción del modelo de regresión lineal

Comenzamos con el modelo de regresión lineal. Inicialmente, una vez divididos el conjunto de datos en entrenamiento y prueba (80, 20 % respectivamente), **realizaremos una primera regresión con todas y cada una de las variables del modelo, incluidas todas las posibles interacciones**. De este modo, aunque no sea el modelo definitivo podremos filtrar aquellas variables más relevantes de cara a facilitar el proceso de selección clásica en lugar de ejecutar directamente la selección con todos los posibles parámetros:

```
formInt<-formulaInteracciones(input_cont,1)
modelo1<-lm(formInt,data=data_train)
# Funcion que muestra tanto el AIC - SBC - R2 train y test (y su diferencia) - Num. parametros
mostrar.estadisticas(modelo1, data_train, data_test, "lm", "varObjCont")
```

```
## Train:  0.7529327 ; Test:  0.6913446 ; Dif. (Train-Test):  0.0615881 ; AIC: 48831.09 ; SBC:  50898.67
## Numero de variables:  304
```

Analizando las estadísticas resultantes, observamos que el R2 obtenido en el conjunto de prueba es significativamente menor que en el conjunto de entrenamiento (diferencia de 0.06 entre ambos), lo que implica un exceso de parámetros y, como consecuencia, **un claro sobreajuste en el modelo**. Con el objetivo de mejorar los resultados, analicemos la importancia de cada una de las variables con respecto al modelo inicial, mediante la función *modelEffectSizes*:

```
variacion.r2 <- modelEffectSizes(modelo1, Print = FALSE)
summary(variacion.r2$Effects[, 4])
```

```
##      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.      NA's
## 0.0000007 0.0000380 0.0000959 0.0002327 0.0002187 0.0065731      1
```

De la salida anterior, debemos destacar el tercer cuartil: de todas las variables del modelo, **el 75 % provocarían una disminución en el R2 de 0.0002 o menos, un valor muy pequeño en comparación con otras variables donde la pérdida sería de 0.006 (valor máximo)**. Es decir, existe un contraste entre variables poco significativas y variables muy significativas. Por ello, dado que la diferencia entre el tercer cuartil y el valor máximo es muy alta, analicemos las variables más atípicas, es decir, las que aportan mayormente al R2:

```
variables.mas.imp <- names(boxplot(variacion.r2$Effects[, 4], plot = FALSE)$out)
```

```
##      [,1]      [,2]
## [1,] "Age_under19_Ptge" "CCAA:PersonasInmueble"
## [2,] "CCAA:Age_0.4_Ptge" "CCAA:Explotaciones"
## [3,] "CCAA:Age_under19_Ptge" "CCAA:logxForeignersPtge"
## [4,] "CCAA:SameComAutonPtge" "CCAA:logxAgricultureUnemploymentPtge"
## [5,] "CCAA:SameComAutonDiffProvPtge" "CCAA:logxtotalEmpresas"
```

```
## [6,] "CCAA:ActividadPpal"      "Age_under19_Ptge:Densidad"
## [7,] "CCAA:PobChange_pct"    ""
```

Salvo el campo *Age_under19_Ptge*, el resto de interacciones parecen ser las más significativas en el modelo original, interactuando especialmente con la Comunidad Autónoma desde el porcentaje de menores de edad (*under_19* y *Age_0_4*) hasta el número de extranjeros o residentes en la misma CCAA o en diferente provincia. Por tanto, de cara a un segundo modelo mantendremos dichas interacciones además de las columnas originales, ya que puede ocurrir que alguna variable sea más significativa sin tener que interactuar con otra:

```
mostrar.estadisticas(modelo1.2, data_train, data_test, "lm", "varObjCont")
```

```
## Train: 0.7391653 ; Test: 0.7281157 ; Dif. (Train-Test): 0.01104957 ; AIC: 48773.34 ; SBC: 49451.24
## Numero de variables: 99
```

Reduciendo el número de parámetros de 304 a 99, el modelo mejora prácticamente en todos los sentidos, tanto un AIC como SBC más bajos, además de recortar la diferencia entre ambos R2 (mejorando en el caso del conjunto de prueba de 0.69 a 0.72).

2.1 Selección de variables clásica

No obstante, el modelo continua teniendo demasiados parámetros, por lo que realizamos una selección clásica empleando este último modelo, mediante los criterios AIC-both, SBC-both, AIC-forward, SBC-forward, AIC-backward y SBC-backward, devolviendo sus resultados en un *dataframe* como sigue a continuación:

	R.2.train	R.2.test	Diferencia	AIC	SBC	N.Parametros
## AIC-both	0.7315348	0.7239069	0.007627932	48872.65	49252.27	55
## SBC-both	0.7259493	0.7226639	0.003285443	48950.42	49140.23	27
## AIC-forward	0.7315348	0.7239069	0.007627932	48872.65	49252.27	55
## SBC-forward	0.7264527	0.7223404	0.004112256	48946.47	49163.40	31
## AIC-backward	0.7385320	0.7273508	0.011181250	48757.10	49326.53	83
## SBC-backward	0.7287939	0.7246366	0.004157332	48888.64	49098.78	30

Analizando la tabla resultante, dado su menor número de parámetros quisiera destacar tanto el modelo 2 como el modelo 6, ya que el resto no ha disminuido lo suficiente en cuanto al número de variables se refiere. En contraste, la diferencia entre el R2 train-test y su menor número de variables da una ligera ventaja al modelo 2, aunque el modelo 6 no solo mejora en cuanto a AIC se refiere, sino incluso que el criterio SBC (que penaliza el número de parámetros), da una mayor ventaja al modelo 6 (49098 frente a 49140). Para confirmar el mejor modelo clásico, realizamos una validación cruzada **con un total de 20 repeticiones, empleando 5 grupos**:

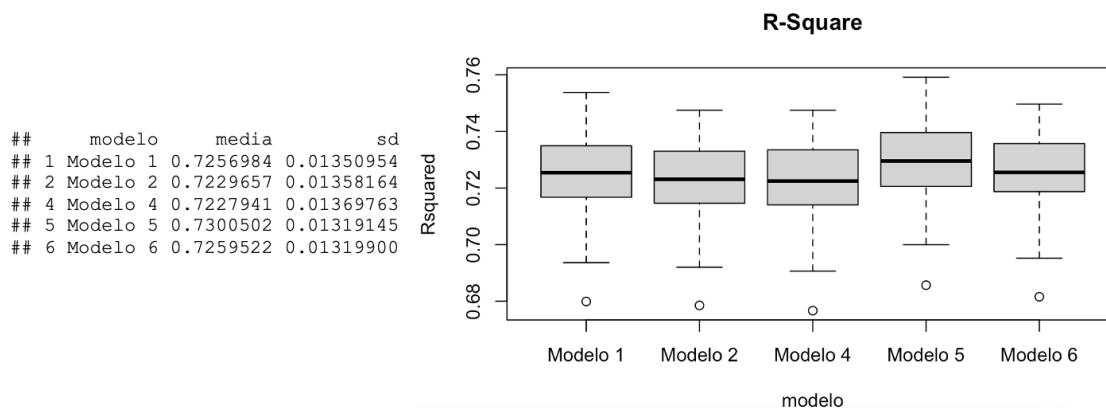


Figure 1: Validación cruzada en los modelos de selección clásicos

Los resultados obtenidos en la validación cruzada arrojan tanto una menor desviación típica (0.0131 frente 0.0135) como una mayor media en los valores R2 del modelo 6 frente al modelo 2 (0.7259 frente 0.7229). Sin embargo, pese a tener tres variables adicionales la mejoría no es muy significativa, además de una amplitud en los diagramas de cajas muy similar. De hecho, ¿Qué variables diferencian a ambos modelos?

```
## ¿Que parámetros tiene el modelo 2 que no tenga el modelo 6?
```

```
##          prop_missings logxAgricultureUnemploymentPtge CCAA:logxForeignersPtge
## dR-sqr          6e-04                      4e-04                      0.0023

## ¿Que parámetros tiene el modelo 6 que no tenga el modelo 2?

##          SameComAutonDiffProvPtge Densidad CCAA:SameComAutonDiffProvPtge
## dR-sqr          2e-04    0.0025                      0.0029
##          Age_under19_Ptge:Densidad
## dR-sqr          0.0028
```

A la vista de los resultados obtenidos, los parámetros exclusivos del modelo 6 presentan una mayor importancia en el R2 con respecto a las variables del modelo 2: a primera vista, **la Densidad y el porcentaje de Población que reside en diferente provincia, así como sus interacciones** parecen ser más relevantes en el modelo que la proporción de *missings*, el porcentaje de desempleo en el sector agrario así como la interacción entre las Comunidades Autónomas con el porcentaje de extranjeros. Bien es cierto que la diferencia entre ambos modelos no es muy significativa: una diferencia en el R2 medio de 0.003 y de 0.0004 en las desviaciones típicas, pese a tener el modelo 6 una interacción adicional. Sin embargo, tanto los criterios AIC como SBC dan una mayor ventaja al modelo 6, además de una mayor importancia en sus variables con respecto al modelo 2, por lo que pueden ser relevantes de cara a una comparación final con los modelos aleatorios. Por tanto, escogemos como modelo candidato al modelo 6 clásico.

2.2 Selección de variables aleatoria

Como última comparación, realizamos una selección aleatoria **realizando 100 repeticiones a partir del 70 % de los datos de entrenamiento (por mayor velocidad)**, con el objetivo de comprobar si existe algún otro modelo que mejore el candidato obtenido en la selección clásica. En primer lugar, analizamos las estadísticas de los **tres mejores modelos aleatorios**:

```
## MODELOS ALEATORIOS

## Modelo aleatorio 1 . Veces que se repite: 2
## Train: 0.7156754 ; Test: 0.7113455 ; Dif. (Train-Test): 0.004329855 ; AIC: 49185.49 ; SBC: 49361.74
## Numero de variables: 25
## Modelo aleatorio 2 . Veces que se repite: 2
## Train: 0.7144549 ; Test: 0.7108354 ; Dif. (Train-Test): 0.003619458 ; AIC: 49209.32 ; SBC: 49372.01
## Numero de variables: 23
## Modelo aleatorio 3 . Veces que se repite: 1
## Train: 0.7140026 ; Test: 0.7111627 ; Dif. (Train-Test): 0.002839889 ; AIC: 49219.6 ; SBC: 49382.29
## Numero de variables: 23
```

En primera instancia, pese a disminuir el número de parámetros, los modelos aleatorios **no mejoran en cuanto a AIC y SBC se refiere**, además de que el valor R2 disminuye ligeramente (de 0.72 en el modelo 6 a 0.71 en los modelos aleatorios). Además, el mejor de los modelos tan solo se repite en dos de las 100 repeticiones aleatorias ¿Y en cuánto a las medias y desviaciones típicas?

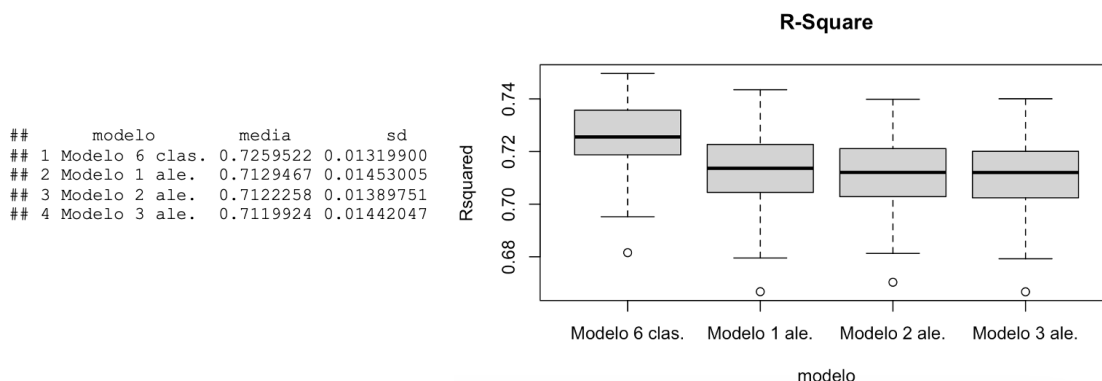


Figure 2: Validación cruzada modelo 6 + modelos de selección aleatorios

2.3 Selección y justificación del modelo ganador

Nuevamente, ninguno de los modelos aleatorios consigue mejorar al modelo 6 en términos de R2 medio y desviación típica, aunque bien es cierto que el segundo modelo aleatorio presenta una desviación típica muy similar (0.013), aunque con una media menor. Por tanto, de todos los modelos evaluados, **el modelo 6 ofrece un mejor resultado tanto en función del criterio AIC, SBC como en desviación típica**. Sin embargo, no podemos declarar el modelo 6 como ganador sin antes hacernos la siguiente pregunta: ¿Existe correlación en sus variables? Uno de los problemas que pudimos analizar en la fase de depuración fue la elevada correlación que presentan muchas de las variables, tanto las edades, el porcentaje de residencia en la misma CCAA como además del total de empresas. Por tanto, debemos eliminar todas aquellas variables, con menor importancia según la salida en *modelEffectSizes*, que presenten una correlación moderada-alta con el resto de parámetros (superior a 0.4 o inferior a -0.4):

modelEffectSizes					Matriz correlacion			
Coefficients								
	SSR	df	pEta-sqr	dR-sqr		Age_under19_Ptge	Age_over65_pct	logxtotalEmpresas
(Intercept)	24398.5990	1	0.0337	NA	Age_under19_Ptge	1.0000000	-0.8644204	0.6558898
CCAA	4853.6283	5	0.0069	0.0019	Age_over65_pct	-0.8644204	1.0000000	-0.6053973
Age_under19_Ptge	404.3579	1	0.0006	0.0002	logxtotalEmpresas	0.6558898	-0.6053973	1.0000000
Age_over65_pct	7533.5386	1	0.0107	0.0029	Construccion	0.4644942	-0.4322092	0.4769535
SameComAutonPtge	7534.2054	1	0.0107	0.0029	logxForeignersPtge	0.4498911	-0.4135373	0.4674308
SameComAutonDiffProvPtge	574.3591	1	0.0008	0.0002				
IndustryUnemploymentPtge	2936.2413	1	0.0042	0.0011				
ServicesUnemploymentPtge	3648.5332	1	0.0052	0.0014	Construccion	0.4644942	0.4498911	0.4769535
Construccion	1012.6068	1	0.0014	0.0004	Age_under19_Ptge	0.4644942	0.4498911	0.4769535
Densidad	6505.8942	2	0.0092	0.0025	Age_over65_pct	-0.4322092	-0.4135373	0.4674308
logxForeignersPtge	2264.8519	1	0.0032	0.0009	logxtotalEmpresas	0.4769535	0.4674308	1.0000000
logxConstructionUnemploymentPtge	2687.7770	1	0.0038	0.0010	Construccion	1.0000000	0.2027565	0.2027565
logxtotalEmpresas	4031.4474	1	0.0057	0.0016	logxForeignersPtge	0.2027565	1.0000000	0.2027565
CCAA:SameComAutonPtge	42173.0923	5	0.0569	0.0164				
CCAA:SameComAutonDiffProvPtge	7399.7126	5	0.0105	0.0029				
Age_under19_Ptge:Densidad	7165.5090	2	0.0101	0.0028				

Figure 3: Salida modelEffectSizes modelo 6 + Matriz de correlación

1. *Age_under19_Ptge* y *Age_over65_pct*: se ha decidido eliminar *Age_under19_Ptge* (junto con *Age_under19_Ptge:Densidad*) dado que *Age_over65_pct* supondría perder prácticamente el mismo R2 sin realizar interacciones con ninguna variable (0.0029 en la columna *dR-sqr*). Por otro lado, el campo *Densidad* pierde significancia al eliminar la interacción con el porcentaje de edad, por lo que lo descartamos.
2. *Age_over65_pct* con *logxtotalEmpresas*, *Construcción* y *logxForeignersPtge*, ya que se perdería el mismo R2 que eliminando tan solo *Age_over65_pct* (0.0029).

Una vez eliminadas las variables, comparamos nuevamente los modelos (modelo 6 modificado como los modelos aleatorios):

##		R2.train	R2.test	AIC	SBC	Num..Parametros
##	Modelo 6 (depurado)	0.7228378	0.7192497	49013.76	49169.67	22
##	Modelo 1 aleatorio	0.7156754	0.7113455	49185.49	49361.74	25
##	Modelo 2 aleatorio	0.7144549	0.7108354	49209.32	49372.01	23
##	Modelo 3 aleatorio	0.7140026	0.7111627	49219.60	49382.29	23

Pese a aumentar tanto el AIC como el criterio SBC, el modelo 6 continua siendo el mejor modelo tanto en términos R2 como incluso por los criterios AIC/SBC, **empleando menos variables que los modelos aleatorios** ¿Y en relación con las medias y desviaciones típicas?

##		Media	sd
##	Modelo 6 (depurado)	0.7206228	0.01349691
##	Modelo 1 aleatorio	0.7129467	0.01453005
##	Modelo 2 aleatorio	0.7122258	0.01389751
##	Modelo 3 aleatorio	0.7119924	0.01442047

Nuevamente, apenas se ha visto afectado, por lo que continua teniendo mayor media y menor desviación. Por tanto, dado su menor número de parámetros (22), así como su mejor bondad de ajuste tanto en términos R2 medio como en desviación típica, **declaramos al modelo 6 depurado como modelo ganador**. Por otro lado, en relación con el modelo de regresión final nos encontramos con la interacción *CCAA:SameComAutonDiffProvPtge* cuyo p-valor sólo es significativo en las regiones de Andalucía y Navarra:

CCAAAN_NA:SameComAutonDiffProvPtge	0.998386	0.119905	8.326	< 2e-16 ***
CCAAAR_CM:SameComAutonDiffProvPtge	0.052157	0.096516	0.540	0.5889

CCAACAT_PV:SameComAutonDiffProvPtge	0.028968	0.082712	0.350	0.7262
CCAACV_EX_AS_BA_CA:SameComAutonDiffProvPtge	0.290994	0.130041	2.238	0.0253 *
CCAAMA_CA_RI_CE_ME_MU_GA:SameComAutonDiffProvPtge	-0.214441	0.179183	-1.197	0.2314

Dado que la interacción solo parece interactuar con una categoría en particular (Andalucía y Navarra) ¿Podríamos eliminar la interacción junto con el campo *SameComAutonDiffProvPtge*? Al hacerlo, el valor R2 en el conjunto de prueba no parece variar significativamente: tan solo disminuye en 0.03 eliminando 6 parámetros del modelo. De hecho, al realizar una nueva validación cruzada no solo disminuye la desviación típica, sino que además la media aumenta ligeramente (0.717 a 0.720), un indicativo de que la interacción no parece relevante. Por lo tanto lo eliminamos:

Eliminando la interaccion => R2-test: 0.7162 ; media: 0.7206228 ; sd: 0.01349691 ; Num. Parámetros: 16
Sin eliminar => R2-test: 0.7192 ; media: 0.7176647 ; sd: 0.01371654 ; Num. Parámetros: 22

Una vez eliminado, obtenemos las nuevas estadísticas del modelo final:

ESTADÍSTICAS DEL MODELO FINAL:

Train: 0.7191562 ; Test: 0.716221 ; Dif. (Train-Test): 0.002935258 ; AIC: 49087.47 ; SBC: 49202.72
Numero de variables: 16 ; sd: 0.01349691

De hecho, y siguiendo el principio de parsimonia, al eliminar la interacción menos relevante **obtenemos una explicación mucho más sencilla al porcentaje de votos**, con menos parámetros. A continuación, analizamos la importancia de las variables mediante la función *summary*:

```
Call:
lm(formula = as.formula(formula.final), data = data_train)

Residuals:
    Min       1Q   Median       3Q      Max
-54.555  -6.081   0.085   6.513  43.823

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    43.14850    2.28942   18.847 < 2e-16 ***
CCAAAN_NA      19.71720    3.85373    5.116 3.20e-07 ***
CCAAAR_CM       5.44023    2.75089    1.978  0.048 *
CCAACAT_PV     19.09879    3.54937    5.381 7.67e-08 ***
CCAACV_EX_AS_BA_CA 14.89312    3.12382    4.768 1.91e-06 ***
CCAAMA_CA_RI_CE_ME_MU_GA 19.73156    3.68024    5.361 8.54e-08 ***
Age_over65_pct    0.13344    0.01513    8.820 < 2e-16 ***
SameComAutonPtge  0.21724    0.02661    8.164 3.87e-16 ***
IndustryUnemploymentPtge -0.08885    0.01314   -6.761 1.49e-11 ***
ServicesUnemploymentPtge -0.04047    0.00567   -7.137 1.06e-12 ***
logxConstructionUnemploymentPtge -0.23632    0.03480   -6.790 1.22e-11 ***
CCAAAN_NA:SameComAutonPtge -0.48867    0.04499  -10.861 < 2e-16 ***
CCAAAR_CM:SameComAutonPtge -0.16526    0.03295   -5.015 5.43e-07 ***
CCAACAT_PV:SameComAutonPtge -0.81445    0.04320  -18.852 < 2e-16 ***
CCAACV_EX_AS_BA_CA:SameComAutonPtge -0.34859    0.03730   -9.345 < 2e-16 ***
CCAAMA_CA_RI_CE_ME_MU_GA:SameComAutonPtge -0.29240    0.04405   -6.637 3.46e-11 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.57 on 6480 degrees of freedom
Multiple R-squared:  0.7192,    Adjusted R-squared:  0.7185
F-statistic: 1106 on 15 and 6480 DF, p-value: < 2.2e-16
```

Analizando la función *summary*, prácticamente todas las variables del modelo son significativas. Para un mayor detalle, analicemos la salida obtenida en *modelEffectSizes*:

```
## lm(formula = as.formula(formula.final), data = data_train)
##
## Coefficients
##
##              SSR df pEta-sqr dR-sqr
## (Intercept)    39680.739  1  0.0520  NA
## CCAA          7258.951  5  0.0099 0.0028
## Age_over65_pct    8690.218  1  0.0119 0.0034
## SameComAutonPtge   7445.415  1  0.0102 0.0029
## IndustryUnemploymentPtge  5107.151  1  0.0070 0.0020
## ServicesUnemploymentPtge  5690.523  1  0.0078 0.0022
## logxConstructionUnemploymentPtge  5151.023  1  0.0071 0.0020
## CCAA:SameComAutonPtge  47449.595  5  0.0615 0.0184
##
```



```
## Sum of squared errors (SSE): 723893.8
## Sum of squared total (SST): 2577567.7
```

Estudiando cada uno de los parámetros, nos encontramos con algunas variables con valores *dR-sqr* muy pequeños. A modo de ejemplo, los campos *SameComAutonPtge* o *Age_over_65_pct* supondrían una pérdida de tan solo 0.0029 y 0.0034 en el R2, respectivamente. No obstante, quisiera remarcar la columna *pEta-sqr*, el cual indica el porcentaje de la varianza explicada por cada variable del modelo. De hecho, **las Comunidades Autónomas, el porcentaje de población superior a 65 años, así como el porcentaje de población que reside en la misma CCAA** son las que mayor cantidad de varianza explican, con más de un 1 % en cada una de ellas (incluso en el caso de *CCAA:SameComAutonPtge* llegando a alcanzar más del 6 % de la varianza explicada). Por tanto, pese a que algunas de las variables supongan una pérdida poco significativa en el R2, el porcentaje de varianza explicada es significativo, por lo que no pueden ser descartadas. Además, los p-valores obtenidos en cada uno de ellos indican que podemos rechazar la hipótesis nula de que los coeficientes no tienen efecto alguno sobre la variable objetivo.

Como conclusión final, a lo largo del proceso de modelado de la regresión lineal, hemos comprobado que ninguna de las variables exclusivas de los modelos 2 y 6 clásicos han resultado de gran utilidad, pese a lo planteado previamente, sino que **el porcentaje de votos a la derecha se ve influido principalmente por las variables comunes a ambos modelos: la CCAA del municipio, el porcentaje de habitantes mayores a 65 años e incluso en función del porcentaje de personas que residen en la misma CCAA, pero especialmente con la interacción entre esta última y la CCAA**, siendo esta última la variable con mayor potencial.

2.4 Interpretación de los coeficientes de dos variables

Finalmente, interpretamos los coeficientes de dos variables obtenidas en el modelo.

1. **CCAACAT_PV** (Cataluña y País Vasco): 19.10. Es decir, **el porcentaje de votos a la derecha aumenta en un 19.10 % aproximadamente si la CCAA a la que pertenece el municipio es Cataluña o País Vasco con respecto a la Comunidad Autónoma de referencia** (Castilla y León).
2. **logxConstructionUnemploymentPtge**: -0.24. Es decir, **por cada incremento unitario en el porcentaje de desempleados en el sector de la construcción, el porcentaje de votos a la derecha se ve reducido en un 0.24 %**. Por tanto, aquellos municipios con mayor porcentaje de paro en la construcción afectan negativamente al voto de la derecha.

3. Construcción del modelo de regresión logística

Una vez construido el modelo de regresión lineal, continuamos con el modelo de regresión logística. En primer lugar, y al igual que en el apartado anterior, elaboramos un primer modelo con todas las variables e interacciones (aunque no se trate del modelo definitivo):

```
formInt.bin<-formulaInteracciones(input_bin, 1)
modelo1.bin<-glm(formInt.bin,data=data_train.bin, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
mostrar.estadisticas(modelo1.bin, data_train.bin, data_test.bin, "glm", "varObjBin")
```

```
## Train: -10.44845 ; Test: -11.01544 ; Dif. (Train-Test): 0.5669854 ; AIC: 99295.52 ; SBC: 101356.3
## Numero de variables: 304
```

Como podemos observar en la salida anterior, con un total de 304 parámetros el modelo no logra converger, principalmente por un motivo: **existen demasiadas variables**, lo cual se traduce en valores pseudo-R2 negativos, es decir, **la inclusión de demasiadas variables está penalizando la calidad del modelo**. Con respecto a las interacciones, debemos recoger únicamente aquellas con mayor relevancia. Por ello, ejecutamos la función *impVariablesLog*:

```
importancia.var <- impVariablesLog(modelo1.bin, "varObjBin", data_train.bin)
```

```
##      Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
## -10.94997 -1.13104 -0.05853 -0.86781  1.29203  4.63291
```

Analizando la salida obtenida en el *summary* anterior, cabe destacar la mediana obtenida (-0.05), es decir, **un 50 % de las variables del modelo presentan una importancia negativa con respecto al modelo de regresión**, es decir, están penalizando las estimaciones obtenidas en el pseudo-R2. Por el contrario, si nos fijamos en el tercer cuartil, un 25 % de las variables son las que aportan la mayor importancia al modelo, **sobresalen** con respecto al resto. Por tanto, de cara a un segundo modelo, y con el fin de reducir el coste computacional en la selección clásica, **filtramos aquellas variables cuya importancia sea mayor al tercer cuartil (1.29)**. El resto de interacciones, que apenas tendrán efecto sobre el modelo las descartamos:

```
variables.mas.imp <- importancia.var[which(importancia.var$V5 > 1.29203), "V2"]
# Ejemplo de algunas de las variables mas importantes (Top 5)
```

```
## [1] "IndustryUnemploymentPtge:ActividadPpal", "Age_over65_pct:Densidad", "CCAA:ForeignersPtge"
## [4] "CCAA:IndustryUnemploymentPtge", "Densidad:prop_missings"
```

```
length(variables.mas.imp)
```

```
## [1] 21
```

A primera vista, nos encontramos con que las variables más importantes corresponden con interacciones (concretamente 21). A diferencia del modelo de regresión lineal, dichas interacciones no solo corresponden con la Comunidad Autónoma, sino incluso con la Densidad o la Actividad Principal del municipio, incluso variables que eran relevantes en el modelo lineal también lo son aquí (Age_over65_pct o IndustryUnemploymentPtge). Una vez recuperadas las interacciones más importantes, elaboramos un segundo modelo junto con las variables originales:

```
modelo1.2.bin<-glm(formInt.bin,data=data_train.bin, family = binomial)
```

```
## Train: 0.4442207 ; Test: 0.4306761 ; Dif. (Train-Test): 0.0135446 ; AIC: 4990.907 ; SBC: 5668.802
## Numero de variables: 100
```

Como podemos observar, no solo hemos conseguido reducir el número de parámetros (de 304 a 100), sino además que los valores obtenidos tanto del conjunto de datos *train* como *test* se corresponden con valores comunes en el pseudoR2 (del orden de 0.4); además de unos criterios AIC y SBC mucho menores, reduciendo de 99.295 a 4.990 en el caso de AIC, por ejemplo.

3.1 Selección de variables clásica

No obstante, pese a que el modelo consigue converger, continúa teniendo demasiados parámetros. Por tanto, partiendo de este último modelo realizamos una selección clásica del mismo modo que en la regresión lineal: empleando los criterios AIC-both, SBC-both, AIC-forward, SBC-forward, AIC-backward y SBC-backward, devolviendo sus resultados en un *dataframe*:

##	R.2.train	R.2.test	Diferencia	AIC	SBC	N.Parametros
## AIC-both	0.4296977	0.4292625	0.0004352367	5008.097	5319.928	46
## SBC-both	0.4095392	0.4118443	-0.0023051484	5113.867	5195.214	12
## AIC-forward	0.4298391	0.4294172	0.0004218693	5008.879	5327.489	47
## SBC-forward	0.4095392	0.4118443	-0.0023051484	5113.867	5195.214	12
## AIC-backward	0.4425879	0.4287903	0.0137976379	4960.982	5489.739	78
## SBC-backward	0.4197849	0.4131002	0.0066846765	5043.548	5185.906	21

Analizando los resultados obtenidos en la selección clásica, **los modelos 1 y 5 (AIC-both y AIC-backward) presentan un criterio AIC menor**, aunque con un elevado número de parámetros, especialmente en el quinto modelo, donde incluso hay una mayor diferencia entre ambos pseudo-R2 (0.01). Por el contrario, los modelos 2 y 6 (SBC-both y SBC-backward) ofrecen un menor número de variables, además de un valor SBC significativamente menor (5195 y 5185, respectivamente). Por tanto, pese al menor valor pseudo-R2 que presentan, los modelos 2 y 6 ofrecen unos resultados muy similares a los modelos 1 y 5, con una diferencia de tan solo 0.01 en el R2, empleando 12 y 21 parámetros, respectivamente. Por tanto, **todo apunta a los modelos 2 y 6 como posibles modelos candidatos**, con valores pseudoR2 muy similares. Sin embargo, llama la atención la diferencia negativa entre los valores *train* y *test* del modelo 2.

Por lo general, en cualquier modelo de aprendizaje automático el conjunto de datos de entrenamiento obtiene un mejor resultado en comparación con los datos de prueba. No obstante, puede ocurrir que los resultados en la validación/prueba sean ligeramente superiores a los datos de entrenamiento, en función del modo en el que se hayan

dividido los datos (valor de la semilla). Dado que la partición ha sido aleatoria, puede ocurrir que el conjunto de entrenamiento sea más difícil de interpretar que los datos de prueba, obteniendo resultados confusos.

Por tanto, los criterios AIC/SBC no resultan suficientes para decidir cual es el mejor modelo clásico, por lo que realizamos una **validación cruzada del mismo modo que en la regresión lineal**. De este modo podremos comprobar si los resultados son independientes en función de la partición de los datos: **si la desviación típica en los valores ROC es baja, significaría que la partición empleada en la selección clásica no sería la más adecuada. En caso contrario, podría tratarse de un posible sobreajuste en el modelo 2:**

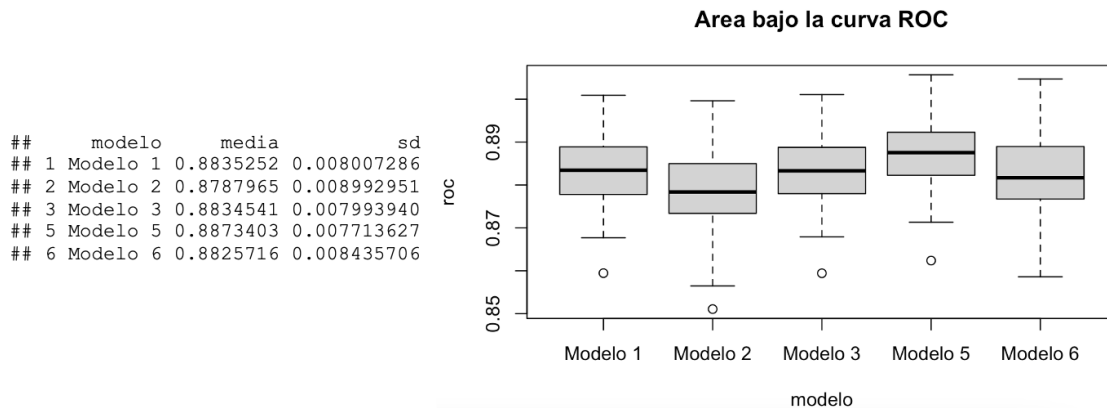


Figure 4: Validación cruzada en los modelos de selección clásicos

Analizando tanto la tabla como los diagramas de caja, observamos que la desviación típica en el modelo 2 no es muy significativa (0.008), por lo que puede que la división empleada en la selección clásica no haya sido la más adecuada. Por lo general, tanto la media como las desviaciones típicas obtenidas son muy similares, con mejor resultado en los modelos 1, 3 y 5 (del orden de 0.007), además de una mejor bondad media en el valor AUC (del orden de 0.88). No obstante, **el mayor número de parámetros que presentan dichos modelos no compensa el aumento del valor medio de 0.87 a 0.88**. Además, aunque en el diagrama de caja y bigotes muestren una menor amplitud, la diferencia con los modelos 2 y 6 no es muy relevante (de 0.85 a 0.87 en el valor mínimo). Como última comparación, analizamos el promedio de los p-valores obtenidos en cada modelo:

```
## Modelo 1
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01102 0.09664 0.19209 0.26612 0.96024
## Modelo 2
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000000 0.0000000 0.0000000 0.0571055 0.0000841 0.6845788
## Modelo 3
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01869 0.11223 0.19670 0.27570 0.95923
## Modelo 5
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01833 0.13422 0.23408 0.38621 0.97194
## Modelo 6
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.000e+00 0.000e+00 1.856e-05 2.128e-02 2.115e-02 1.551e-01
```

Pese a tener mejores medias y desviaciones típicas, **muchas de las variables en los modelos 1, 3 y 5 no son significativas**. A modo de ejemplo, la mediana en cada uno de ellos indica que un 50 % de los coeficientes no disminuye su p-valor de 0.05 (poca significancia), por lo que los descartamos. En relación con el modelo 6, ha demostrado tener, en la validación cruzada, una media superior al modelo 2 (0.88 frente a 0.87). No obstante, la diferencia entre ambos es muy pequeña. A modo de ejemplo, añadiendo 9 parámetros más, el modelo 6 solo consigue una mejora en el pseudoR2 (test) de tan solo 0.02, además de una diferencia en el AUC de 0.01. Por otro lado, si nos fijamos en la siguiente salida, **mientras que el modelo 6 emplea interacciones, el modelo 2 obtiene unos resultados similares sin emplear ninguna interacción:**

```
## Modelo 6: CCAA + Age_over65_pct + ForeignersPtge + AgricultureUnemploymentPtge + IndustryUnemploymentPtge +
## ActividadPpal + PobChange_pct + prop_missings + ForeignersPtge:ActividadPpal + CCAA:IndustryUnemploymentPtge

## Modelo 2: CCAA + ForeignersPtge + AgricultureUnemploymentPtge + prop_missings + Age_over65_pct + PobChange_pct + SUPERFICIE
```

Además, en cuanto a los p-valores, las variables del segundo modelo son mucho más relevantes que en el sexto (tercer cuartil = $8e-05$ frente a 0.021), es decir, **las variables del modelo 2 son generalmente más significativas que las del modelo 6**. Por tanto, de cara a una comparación final con los modelos aleatorios, dado el buen valor pseudoR2 y AUC obtenido en proporción con un menor número de parámetros **escogemos como modelo candidato al modelo 2**, pese a la diferencia negativa que presenta entre los valores *train* y *test*.

3.2 Selección de variables aleatoria

Una vez realizada la selección clásica, elaboramos los modelos aleatorios a partir de la combinación de **todas las variables y sus interacciones**, un proceso computacionalmente más costoso pero que permite comprobar si hemos omitido alguna interacción relevante en los primeros pasos:

```
## MODELO 2 CLASICO

## Train: 0.4095392 ; Test: 0.4118443 ; Dif. (Train-Test): -0.002305148 ; AIC: 5113.867 ; SBC: 5195.214
## Numero de variables: 12

## MODELOS ALEATORIOS (TOP 3)

## Modelo aleatorio 1 . Veces que se repite: 5
## Train: 0.4092791 ; Test: 0.4070475 ; Dif. (Train-Test): 0.002231567 ; AIC: 5116.109 ; SBC: 5197.456
## Numero de variables: 12
## Modelo aleatorio 2 . Veces que se repite: 4
## Train: 0.4046166 ; Test: 0.4083817 ; Dif. (Train-Test): -0.003765128 ; AIC: 5150.301 ; SBC: 5211.312
## Numero de variables: 9
## Modelo aleatorio 3 . Veces que se repite: 3
## Train: 0.4078011 ; Test: 0.4075987 ; Dif. (Train-Test): 0.000202435 ; AIC: 5126.85 ; SBC: 5201.418
## Numero de variables: 11
```

3.3 Selección y justificación del modelo ganador

Analizando el top 3 modelos aleatorios, **llama la atención el primer modelo más repetido (en 5 ocasiones)**, con el mismo número de variables que el modelo 2, además de unos valores pseudo-R2 muy similares entre el train y test (0.40 en ambos). No obstante, a simple vista no parece estar claro cual puede ser el modelo ganador, pero si realizamos la validación cruzada:

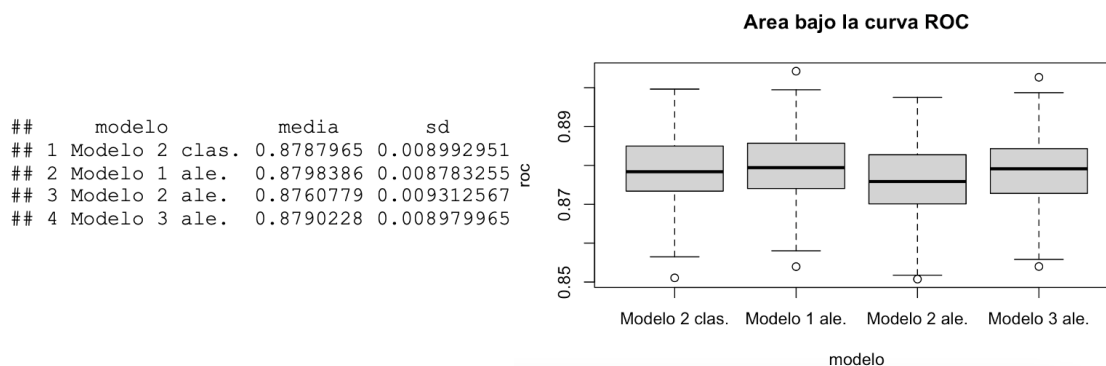


Figure 5: Validación cruzada modelo 2 + modelos de selección aleatorios

En ambos casos, tanto los valores medios del área bajo la curva ROC como la desviación típica son muy similares (bondad media muy parecida, en torno a 0.87). En el caso del segundo modelo aleatorio, pese a tener tan solo nueve parámetros, los valores AIC y SBC son los más altos, además de obtener una desviación típica ligeramente superior al resto de variables, por lo que es posible que el modelo esté omitiendo variables relevantes. Para comprobarlo, analicemos los modelos obtenidos:

```
## [1] "Age_over65_pct+AgricultureUnemploymentPtge+CCAA+Densidad+ForeignersPtge+PobChange_pct"
## [2] "AgricultureUnemploymentPtge+CCAA+ForeignersPtge+prop_missings"
## [3] "Age_over65_pct+AgricultureUnemploymentPtge+CCAA+Densidad+ForeignersPtge"
```

Llama la atención que el modelo 3 aleatorio tan solo se diferencie del primero por el campo *PobChange_pct*. Al eliminar el campo, tanto el AIC como el SBC apenas aumentan en unas unidades (de 5116 a 5126 en el AIC, de 5197 a 5201 en el SBC). Por otro lado, ambos modelos son muy similares en cuanto a valor medio y desviación típica se refiere (0.87 y 0.008, respectivamente) con apenas unas milésimas de diferencia: **el porcentaje de cambio en la población no parece ser un factor significativo en el voto a la derecha**. Por tanto, de los modelos aleatorios, el modelo 3 parece ser un modelo candidato a comparar con el segundo modelo clásico. Para ello, en este último modelo se han encontrado los campos *Age_over65_pct* y *PobChange_pct*, variables con elevada correlación (-0.52), por lo que eliminamos el menos relevante (en función de la salida en *impVariablesLog*):

```
impVariablesLog(estadisticas.modelos.bin[2]$SBC~both`, "varObjBin", data_train.bin)
impVariablesLog(glm(tercer.modelo.aleatorio, "varObjBin", data_train.bin))
```

MODELO 2 CLASICO	IMPORTANCIA	MODELO 3 ALEATORIO	IMPORTANCIA
SUPERFICIE	0.00147	Age_over65_pct	0.00233
PobChange_pct	0.00152	AgricultureUnemploymentPtge	0.00242
AgricultureUnemploymentPtge	0.00228	Densidad	0.00549
Age_over65_pct	0.00299	ForeignersPtge	0.00838
prop_missings	0.00450	CCAA	0.32887
ForeignersPtge	0.00892		
CCAA	0.32638		

```
## Correlacion entre Age_over65_pct y PobChange_pct: -0.5292537
```

```
## Modelo 2 clásico (modificado):
```

```
## Train: 0.408012 ; Test: 0.4126571 ; Dif. (Train-Test): -0.004645077 ; AIC: 5125.032 ; SBC: 5199.6
## Numero de variables: 11
```

```
## Modelo 3 aleatorio:
```

```
## Train: 0.4078011 ; Test: 0.4075987 ; Dif. (Train-Test): 0.000202435 ; AIC: 5126.85 ; SBC: 5201.418
## Numero de variables: 11
```

Pese a eliminar el campo *PobChangePct*, los criterios AIC y SBC continúan siendo muy similares entre si, con una diferencia muy pequeña. Sin embargo ¿Cómo ha variado la desviación típica y la media?

```
##          modelo      media      sd
## 1 Modelo 2 (modificado) 0.8775700 0.009154448
## 2   Modelo 3 aleatorio 0.8790228 0.008979965
```

Nuevamente, los valores son muy similares, aunque con una ligera ventaja en el modelo 3 aleatorio. Si además nos fijamos en la importancia de las variables mostrada anteriormente, el modelo 3 aleatorio **emplea dos variables originales menos que en el modelo 2**. Por un lado, ambos modelos **parecen coincidir en la importancia de la Comunidad Autónoma, el porcentaje de extranjeros, la población mayor a 65 años, así como el desempleo en el sector agrario**, especialmente en la CCAA, con mayor peso en ambos casos. No obstante, mientras que el modelo 2 clásico emplea la superficie, la proporción de valores atípicos (*prop_missings*) y el porcentaje de cambio en la población, el tercer modelo aleatorio sustituye todos ellos por el campo *Densidad*, **aportando prácticamente la misma importancia que la suma de los tres anteriores (0.005)**. Si a lo anterior le añadimos una menor desviación típica (0.008 frente a 0.009), un valor medio ligeramente superior (0.879 frente a 0.877), una diferencia entre los valores pseudo-R2 menor y positiva, además del mismo principio de parsimonia: *ante dos explicaciones, la más sencilla suele ser la más probable*. Por tanto, dado que el tercer modelo aleatorio emplea un menor número de variables originales (5), **lo consideramos como modelo ganador**:

ESTADISTICAS DEL MODELO FINAL:

```
Train: 0.4078011 ; Test: 0.4075987 ; Dif. (Train-Test): 0.000202435 ; AIC: 5126.85 ; SBC: 5201.418
Numero de variables: 11 ; sd: 0.008979965
```

A continuación, se ha planteado la posibilidad de **añadir algunas de las interacciones más significativas planteadas inicialmente**, con el propósito de comprobar si el modelo mejora:

Interaccion	AIC	SBC	PseudoR2 train	PseudoR2 test	sd
-------------	-----	-----	----------------	---------------	----

1	IndustryUnemploymentPtge:ActividadPpal	5119.72	5214.62	0.40932	0.41076	0.00890
2	Age_over65_pct:Densidad	5130.51	5218.64	0.40784	0.40803	0.00899
3	CCAA:ForeignersPtge	5121.14	5229.6	0.40962	0.41048	0.00903
4	CCAA:IndustryUnemploymentPtge	5083.37	5198.61	0.41423	0.40708	0.00845
5	Densidad:prop_missings	5124.47	5219.38	0.40877	0.41081	0.00892
6	Sin interacciones	5126.85	5201.41	0.40780	0.40759	0.00897

Incluso añadiendo algunas de las interacciones más relevantes, en la mayoría de los casos se obtiene un valor pseudoR2 en el conjunto *test* ligeramente superior con respecto a los datos de entrenamiento (diferencia negativa). Incluso en aquellas interacciones donde la diferencia es positiva (CCAA:IndustryUnemploymentPtge), pese a que el pseudo-R2 aumente en el conjunto de entrenamiento (0.40 a 0.41), en el caso de los datos de prueba parece verse reducido ligeramente (de 0.4075 a 0.4070), por lo que aparentemente **las interacciones no aportan mejoría al modelo de regresión logística**.

A continuación, analizamos los coeficientes del modelo ganador mediante la función *summary*:

```
Call:
glm(formula = formula.final.bin.aleatorio, family = binomial,
    data = data_train.bin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7761  -0.1598   0.3932   0.6425   3.4267

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.679445    0.168185   9.986 < 2e-16 ***
Age_over65_pct  0.016838    0.003780   4.454 8.42e-06 ***
AgricultureUnemploymentPtge -0.016630    0.003613  -4.603 4.16e-06 ***
CCAAAN_NA    -3.357578    0.127897 -26.252 < 2e-16 ***
CCAAAR_CM    -1.572530    0.107108 -14.682 < 2e-16 ***
CCAAAT_PV    -7.715219    0.398165 -19.377 < 2e-16 ***
CCAAEX_EX_AS_BA_CA -1.822037    0.118062 -15.433 < 2e-16 ***
CCAAAR_MA_CA_RI_CE_ME_MU_GA -0.104502    0.169245  -0.617    0.537
DensidadBaja  0.540913    0.118780   4.554 5.27e-06 ***
DensidadAlta  1.022174    0.175629   5.820 5.88e-09 ***
ForeignersPtge  0.050303    0.006158   8.169 3.11e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8620.2  on 6495  degrees of freedom
Residual deviance: 5104.8  on 6485  degrees of freedom
AIC: 5126.8

Number of Fisher Scoring iterations: 7
```

Analizando los coeficientes, llama la atención las provincias de Madrid, Cantabria, Rioja, Ceuta, Melilla, Murcia y Galicia (MA_CA_RI_CE_ME_MU_GA), cuyo p-valor corresponde con el máximo de todo el modelo (0.537). Al tratarse de la categoría con menor número de variables (816), puede ocurrir que tenga una menor representación con respecto al resto de Comunidades. No obstante, **no se ha optado por agrupar** la categoría con la región más cercana en cuanto a media (AR_CM), dado que los criterios de error AIC/SBC aumentan considerablemente, además de disminuir el valor pseudoR2 de 0.40 a 0.39, un indicativo de que estamos perdiendo información relevante al agrupar ambas categorías.

```
# Sin agrupar MA_CA_RI_CE_ME_MU_GA
Train: 0.4078011 ; Test: 0.4075987 ; AIC: 5126.85 ; SBC: 5201.418
# Agrupando AR_CM con MA_CA_RI_CE_ME_MU_GA = AR_CM_MA_CA_RI_CE_ME_MU_GA
Train: 0.3950895 ; Test: 0.3908024 ; AIC: 5234.426 ; SBC: 5302.216
```

3.4 Selección del punto de corte óptimo

Una vez elegido el modelo final, debemos **evaluar cual es el punto de corte que ofrece un mejor resultado**. Para ello, obtenemos tanto el punto de corte que maximice la tasa de aciertos como el índice de Youden, generando una rejilla con todos los posibles puntos de corte (de 0 a 1 en intervalos de 0.01):

```
## Índice que maximiza la tasa de aciertos: 0.54
## Índice de Youden: 0.59
```

Una vez obtenidos ambos puntos de corte, comparamos los estadísticos obtenidos:

```
sensEspCorte(modelo.final.bin.aleatorio,data_test.bin,"varObjBin",0.54,"1") # Max. tasa aciertos
```

```
##      Accuracy      Sensitivity      Specificity Pos Pred Value Neg Pred Value
##      0.8367221      0.9404762      0.6666667      0.8222029      0.8723404
```

```
sensEspCorte(modelo.final.bin.aleatorio,data_test.bin,"varObjBin",0.59,"1") # Indice Youden
```

```
##      Accuracy      Sensitivity      Specificity Pos Pred Value Neg Pred Value
##      0.8305607      0.9057540      0.7073171      0.8353156      0.8207547
```

Analizando los porcentajes obtenidos, en ambos índices se obtiene una precisión muy parecida (en torno al 83 %). No obstante, mediante el primer índice se consigue maximizar la sensibilidad o tasa de verdaderos positivos, es decir, de cada 100 municipios en los que el modelo considera que hay un mayor número de votos a la derecha, 94 de ellos son verdaderos positivos. Por otro lado, no solo es capaz de maximizar la tasa de sensibilidad, sino además el valor predictivo negativo: de cada 100 municipios en los que el modelo considera que no hay mayor número de votos a la derecha, ha predicho correctamente 87 de ellos (reduciendo el número de falsos negativos). Sin embargo, el objetivo de un modelo no es solo conseguir que sea capaz de acertar en qué municipios resulta ganador la derecha, sino además **ser capaz de acertar también (en la medida de lo posible) aquellos municipios en los que no**, esto es, **buscar un equilibrio entre los porcentajes de sensibilidad y especificidad**.

A modo de ejemplo, el primer índice es capaz de acertar con alta precisión qué municipios votan a la derecha, mejorando tanto el porcentaje de verdaderos positivos (94 %) como el valor predictivo negativo (87 %). No obstante, la menor sensibilidad que presenta (66 %) refleja que el modelo “pasa por alto” un elevado número de municipios que deberían considerarse 0 (no gana la derecha), pero que los está clasificando como municipios ganadores (1): **augmenta la probabilidad de obtener falsos positivos**. Por el contrario, el índice de Youden, aunque con una sensibilidad y un valor predictivo negativo menor, consigue un mayor equilibrio entre los porcentajes de sensibilidad y especificidad, ya que pese a reducir la sensibilidad, no solo continúa situándose por encima del 90 %, sino además consigue mejorar la especificidad en un 70 %. Por tanto, con el objetivo de lograr un mayor balance entre sensibilidad y especificidad, **escogemos el índice de Youden como punto de corte óptimo**.

Tras elegir el punto de corte óptimo, observamos que las medidas de clasificación son muy similares en ambos conjuntos de datos (entrenamiento y prueba), con porcentajes ligeramente superiores en el entrenamiento.

```
sensEspCorte(modelo.final.bin.aleatorio,data_train.bin,"varObjBin",0.59,"1")
```

```
##      Accuracy      Sensitivity      Specificity Pos Pred Value Neg Pred Value
##      0.8348214      0.9090458      0.7131247      0.8385917      0.8270500
```

```
sensEspCorte(modelo.final.bin.aleatorio,data_test.bin,"varObjBin",0.59,"1")
```

```
##      Accuracy      Sensitivity      Specificity Pos Pred Value Neg Pred Value
##      0.8305607      0.9057540      0.7073171      0.8353156      0.8207547
```

En relación con el área bajo la curva ROC, los valores tanto en el entrenamiento como prueba son muy similares, con porcentajes muy cercanos al 90 % en ambos casos:

```
## Train AUC: 0.8799913 ; Test AUC: 0.8769293
```

3.5 Interpretación de los coeficientes de dos variables

Por último, interpretamos los coeficientes de dos variables incluidas en el modelo:

1. **DensidadAlta**: 1.022. Es decir, el ODD de que en un municipio, cuya densidad de población es alta (> 5 hab./ha), resulte ganador la derecha es $e^{1.022} = 2.78$ veces mayor que el ODD de un municipio cuya densidad es Muy Baja (< 1 hab./ha), esto es, la categoría de referencia. Por tanto, y dado que el coeficiente de *DensidadBaja* es de 0.54, podemos decir que **conforme aumenta la densidad en el municipio, el ODD de que resulte ganador la derecha también aumenta**.
2. **AgricultureUnemploymentPtge**: -0.016. Es decir, por cada unidad en la que se decremента el porcentaje de desempleados en el sector agrario, la ODD de que en el municipio resulte ganador la derecha aumenta en un $e^{-0.016} = 0.98$ %. Por tanto, cuanto mayor sea el porcentaje de desempleo en el sector agrario, el ODD de que resulte ganador la derecha disminuye.