Miguel Angel Flores Torres

Diciembre, 20201.

máster en big data y business analytics

Minería de datos y Modelización Predictiva

# 1. Introducción al objetivo del problema

El objetivo de este trabajo es estimar, a través de modelos de regresión lineal el porcentaje de votos de los partidos distintos del PP, Ciudadanos, PSOE y Podemos en cada uno de los municipios de España,.

Asimismo, a través de modelos de regresión logística, se estimará para cada uno de los municipios, si el porcentaje de abstención es superior a un 30%, en este caso, la variable objetivo tomará el valor de 1 el % de abstención es superior al 30% y 0 si este es menor.

Las variables objetivo anteriormente descritas serán estimadas a partir de otra serie de variables explicativas como son el código de provincia, la comunidad autónoma a la que pertenece cada uno de los municipios, Actividad principal y densidad del municipio las cuales serán consideradas como variables cualitativas, mientras que el porcentaje de ciudadanos por rango de edad, porcentaje de mujeres, porcentaje de extranjeros, porcentaje de ciudadanos que residen en la misma Comunidad Autónoma en la que nacieron, porcentaje de ciudadanos que reside en la misma CCAA en la que nacieron, pero distinta provincia, porcentaje de ciudadanos que reside en la distinta CCAA de la que nacieron, el porcentaje de parados por rango de edad o tipo de sector, entre otras, han sido consideradas como variables cuantitativas a considerar en los modelos de regresión.

# 2. Importación, análisis y tratamiento de los datos

## 2.1 Asignación de los tipos de variable

Una vez que han sido importado la base de datos, se ha procedido al análisis del conjunto de datos en donde se han asignado el tipo de variable para cada una de las variables explicativas. El primer paso ha sido definir las variables cualitativas como tipo factor para que estas puedan ser incluidas dentro de nuestros modelos.

datos[,c(1,3,7,11:12,34, 38)] <- lapply(datos[,c(1,3,7,11:12,34, 38)], factor)

Como siguiente paso, se han definido la variable objetivo binaria y variable objetivo continua y se ha procedido a la eliminación del conjunto de datos para su posterior análisis.

Es importante destacar que la variable nombre ha sido considerada como variable identificativa y por lo tanto esta también ha sido eliminada de nuestro conjunto de datos.

varObjCont<-datos$Otros\_Pct  
varObjBin<-datos$AbstencionAlta  
datosfinal<-as.data.frame(datos[,-c(1,7,10)])

## 2.2 Corrección de los errores detectados en el conjunto de datos

Después de haber definido las variables objetivo, se ha procedido al análisis de nuestro conjunto de datos, para determinar los valores erróneos dentro del mismo y se ha encontrado lo siguiente:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Número observaciones** | **Tipo de error** |
| Densidad | 92 | valores mal codificados |
| Explotaciones | 189 | observaciones cuyo valor es 99,999 y han sido consideradas como valor missing |
| ForeignersPtge | 653 | observaciones cuyo valor es menor que 0 y por tanto han sido consideradas como valores fuera de rango |
| SameComAutonPtge | 3 | observaciones cuyo valor es mayor que 100 y estas han sido consideradas como valores fuera de rango |
| totalEmpresas | 5 | valores missing |
| Industria | 188 | valores missing |
| Construccion | 139 | valores missing |
| ComercTTEHosteleria | 9 | valores missing |
| Servicios | 62 | valores missing |
| inmuebles | 138 | valores missing |

Para las variables de Densidad, Explotaciones, ForeignerPtge y SameComAutonPtge los valores que han sido encontrados como mal codificados o fuera de rango, han sido trasformados a valores missing para su futura imputación.

#Cambio a NAs  
datosfinal$Densidad<-recode.na(datosfinal$Densidad,"?")

## Recoded 92 values to NA.

datosfinal$Explotaciones<-replace(datosfinal$Explotaciones,  
 which(datosfinal$Explotaciones==99999),NA)  
datosfinal$ForeignersPtge<-replace(datosfinal$ForeignersPtge,  
 which(datosfinal$ForeignersPtge<0),NA)  
datosfinal$SameComAutonPtge<-replace(datosfinal$SameComAutonPtge,  
 which(datosfinal$SameComAutonPtge>100),NA)

## 2.3 Análisis de valores atípicos, perdidos e imputaciones

Una vez que hemos identificado los errores detectados en nuestro conjunto de datos y se han transformado en valores perdidos, se ha procedido a analizar si el porcentaje de datos *missing* para cada una de las variables/observaciones con el objetivo de determinar si el porcentaje de datos *missing* es superior al 50%.

datosfinal$prop\_missings<-apply(is.na(datosfinal),1,mean)  
(prop\_missingsVars<-apply(is.na(datosfinal),2,mean))

Una vez aplicada la formula para determinar la proporción de *missings*, se ha encontrado que adicionalmente a las variables descritas en el apartado anterior, las variables SUPERFICIE, PersonasInmueble, Pob2010 y PobChange\_pct también contienen valores perdidos, sin embargo, para ninguna de estas variables, el porcentaje de valores perdidos es superior al 50% y por tanto, se han mantenido todas las variables iniciales para continuar con nuestro estudio.

#Proporción atípicos  
  
atipicos <-(numAtipicos<-sapply(Filter(is.numeric, datosfinal),  
 function(x) atipicosAmissing(x)[[2]])/nrow(datosfinal))

describe(Filter(is.numeric, datosfinal))

Adicionalmente, se ha procedido a analizar los datos atípicos de nuestro conjunto de datos, sin embargo, hay variables como los porcentaje de desempleo por sector en donde perfectamente puede haber municipios en donde haya un alto porcentaje de desempleo en los sectores incluidos en nuestro conjunto de datos y por tanto estas observaciones no se han considerado como valores atípicos.

Por otro lado, también se han encontrado datos atípicos en las variables de Población o TotalCensus, sin embargo, estos valores se observan en Comunidades autónomas como Madrid, Andalucía o Cataluña que incluyen ciudades como Madrid, Barcelona o Sevilla donde se concentra una gran parte de la población, por lo que se ha tomado la decisión de no considerarse esta información como valores atípicos.

Por último, para variables como totalEmpresas, ComercTTEHosteleria y Servicios, debido a que son variables con demasiada variabilidad de sus datos estos no se han considerado como valores atípicos y se ha tomado la decisión de categorizar estas variables.

En relación a la imputación de los valores NA de nuestro conjunto de datos en las variables cuantitativos, esta se ha hecho a través de su mediana, mientras que para las variables cualitativos, esta se ha realizado a través del modo aleatorio.

#Imputacion variables continuas  
datosfinal[,as.vector(which(sapply(datosfinal, class)=="numeric"))]<-sapply(  
 Filter(is.numeric, datosfinal),function(x) ImputacionCuant(x,"mediana"))  
  
#Imputación variables categoricas  
  
datosfinal[,as.vector(which(sapply(datosfinal, class)=="factor"))]<-sapply(  
 Filter(is.factor, datosfinal),function(x) ImputacionCuali(x,"aleatorio"))  
  
datosfinal[,as.vector(which(sapply(datosfinal, class)=="character"))] <- lapply(datosfinal[,as.vector(which(sapply(datosfinal, class)=="character"))] ,  
 factor)

## 2.4 Recategorización de variables

Ya que tenemos lista nuestra base de datos, para finalizar, se ha realizado es identificar aquellas variables que podrían re categorizarse por la naturaleza de sus datos. Las variables que se han recategorizado son las siguientes:

* **CCAA** -> Debido a que la frecuencia de las comunidades de Ceuta, Melilla y Murcia tiene una poca representación en dicha variable, estas se han recategoriza do generando una categoría llamada Otros

datosfinal$CCAA<-recode(datosfinal$CCAA,"c('Ceuta','Melilla','Murcia')='Otros'")

* **Industria, Construcción, ComercTTEHosteleria y Servicios** -> Debido a la variabilidad de las observaciones y la poca representación, estas variables se han recategorizado convirtiéndose en variables categóricas en donde se ha asignado un valor de 1 si el municipio cuenta con este tipo cuenta con empresas de este sector y 0 en otro caso

datosfinal$Industria<- ifelse(datosfinal$Industria !=0,1,0)  
datosfinal$Construccion <-ifelse(datosfinal$Construccion != 0, 1,0)  
datosfinal$ComercTTEHosteleria <- ifelse(datosfinal$ComercTTEHosteleria != 0,1,0)  
datosfinal$Servicios <- ifelse(datosfinal$Servicios !=0,1,0)  
datosfinal[,c(27:30)] <- lapply(datosfinal[,c(27:30)], factor)

* **totalEmpresas** -> Similar que el caso anterior, para esta variable se han creado categorías por tramo de numero de empresas y esta variable se ha convertido a factor

datosfinal$totalEmpresas<- ifelse(datosfinal$totalEmpresas == 0, datosfinal$totalEmpresas,  
 ifelse(datosfinal$totalEmpresas >0 & datosfinal$totalEmpresas<=10,10,  
 ifelse(datosfinal$totalEmpresas >10 & datosfinal$totalEmpresas<=20,20,  
 ifelse(datosfinal$totalEmpresas >20 & datosfinal$totalEmpresas<=30,30,  
 ifelse(datosfinal$totalEmpresas >30 & datosfinal$totalEmpresas<=50,50,  
 ifelse(datosfinal$totalEmpresas >50 & datosfinal$totalEmpresas<=100,100,  
 ifelse(datosfinal$totalEmpresas >100 & datosfinal$totalEmpresas<=500,500,1000)))))))  
  
datosfinal[,c(26)] <- factor(datosfinal[,c(26)])  
  
datosfinal$totalEmpresas<-recode(datosfinal$totalEmpresas, "0='0' ;10='1-10';  
 20 ='11-20';30='21-30';50='31 -50';  
 100 ='51-100';500='>100';1000='>100'")

* **ActividadPpal** -> Debido a la poca representatividad de las categorías Construcción e Industría, estas han sido unidas a la categoría Otro

datosfinal$ActividadPpal <- recode(datosfinal$ActividadPpal, "c('Construccion','Industria')='Otro'")

# 3. Transformación de variables explicativas

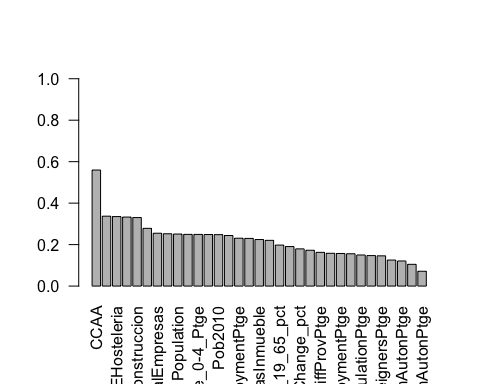
Una vez que se ha realizado la depuración del conjunto de datos a utilizar en nuestros modelos de regresión, se ha procedido a buscar la mejor transformación de nuestras variables cuantitativas que maximice la relación entre nuestras variables objetivo continua y las variables explicativas.

Para nuestro conjunto de datos, las variables explicativas que maximizan su relación con la variable objetivo una vez transformadas son TotalCensus, ForeignersPtge, SameComAutonPtge, Unemploy25\_40\_Ptge, UnemployMore40\_Ptge, IndustryUnemploymentPtge, inmuebles, Pob2010, SUPERFICIE.

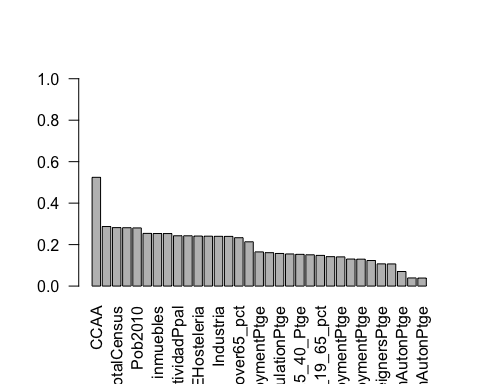
Por otro lado, en relación a la variable objetivo binaria, la mejor transformación para maximizar la relación entre nuestras variables explicativas y la variable objetivo es únicamente la estandarización de las variables cuantitativas.

input\_cont<-cbind(input,Transf\_Auto(Filter(is.numeric, input),varObjCont))  
input\_bin<-cbind(input,Transf\_Auto(Filter(is.numeric, input),varObjBin))

# 4. Detección de las relaciones entre las variables input y objetivo



Como se puede observar en el gráfico de V Cramer, para la variable objetivo continua, las variables explicativas Comunidad Autónoma, Servicios, ComercTTEHosteleria, Industria, Construcción y ActividadPpal son las variables que mas nos ayudarán a predecir la variable objetivo en el modelo de regresión lineal.



Respecto a la variable objetivo binaria, nuevamente la Comunidad Autónoma es la que tiene un gran poder de predicción, seguida por las variables totalEmpresas, TotalCensus, Population y Pob2010.

# 5. Modelo de regresión lineal

## 5.1 Conjunto de datos training y test

Una vez que hemos identificado la mejor transformación de las variables explicativas y las variables que mas nos ayudarán a la predicción a la hora de la construcción de nuestro modelo de regresión lineal, se ha procedido a la creación del conjunto de datos *training* y *test*, los cuales serán utilizados para la determinación del mejor modelo de predicción de nuestra variable objetivo continua.

Para ello, se ha seleccionado una muestra de datos correspondiente al 80% de nuestro conjunto de datos el cual utilizaremos para creación de nuestro modelo, mientras que el resto de datos, serán utilizados para comprobar que el modelo que hemos generado con los datos de entrenamiento funciona con otro conjunto de datos.

Para nuestra construcción del modelo de regresión lineal, nuestro base de datos ha sido repartido entre 6.496 observaciones para el conjunto de datos de entrenamiento y 1.623 observaciones para el conjunto de datos de validación.

todo<-data.frame(input\_cont,varObjCont)  
trainIndex <- createDataPartition(todo$varObjCont, p=0.8, list=FALSE)  
data\_train\_t <- todo[trainIndex,]  
data\_test\_t <- todo[-trainIndex,]  
data\_train\_st <- data\_train\_t[,c(1:33,51)]  
data\_test\_st <- data\_test\_t[,c(1:33,51)]

## 5.2 Selección de variables clásica

Ya que hemos definido el conjunto de datos que utilizaremos de entrenamiento, como siguiente paso para la construcción del modelo de regresión lineal a través de la selección de variables clásica, como primer paso se ha definido el modelo mínimo el cual no contiene ninguna variable y el modelo completo, el cual contiene todas las variables explicativas para la predicción.

Dentro de la selección de variables clásica se han construido 14 modelos, cuyos resultados principales se muestran a continuación:

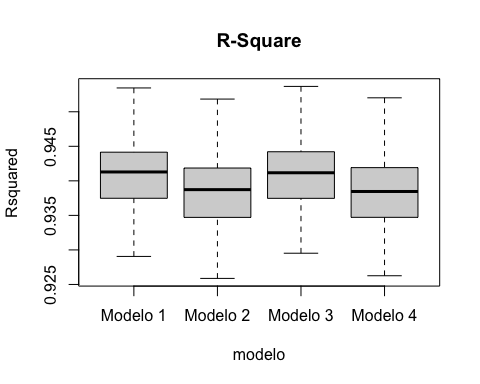
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Tipo** | **Tipo de variables** | **train** | **# variables** |
| Mod1Step | AIC | Transformadas | 0.9424817 | 52 |
| Mod2Step | BIC | Transformadas | 0.9415632 | 33 |
| Mod3Setp | AIC | Originales | 0.9394626 | 42 |
| Mod4Setp | BIC | Originales | 0.9389497 | 27 |
| Mod5Forw | AIC | Transformadas | 0.9424872 | 54 |
| Mod6Forw | BIC | Transformadas | 0.9416374 | 34 |
| Mod7Forw | AIC | Originales | 0.9394626 | 42 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Tipo** | **Tipo de variables** | **train** | **# variables** |
| Mod8Forw | BIC | Originales | 0.9389497 | 27 |
| Mod9Back | AIC | Transformadas | 0.9425139 | 55 |
| Mod10Back | BIC | Transformadas | 0.9417664 | 35 |
| Mod11Back | AIC | Originales | 0.9396926 | 46 |
| Mod12Back | BIC | Originales | 0.9391035 | 29 |
| Mod13Step | AIC | Transformadas e interacciones | 0.9722589 | 306 |
| Mod14Step | BIC | Transformadas e interacciones | 0.9669801 | 91 |

Como se puede observar en la tabla anterior, la de los modelos de predicción construidos bajo la selección de variables clásica son muy similares. A pesar de que en los modelos en donde se han incluido las interacciones son las que mejor predicen nuestra variable objetivo, el número de variables utilizadas en cada uno de los modelos es muy significativa lo que hace que esto penalice a la hora de elegir los modelos candidatos para realizar la validación cruzada de los mismos. Siguiendo el principio de parsimonia, elegiremos los modelos con el menor número de variables utilizados para predecir nuestra variable objetivo los cuales se indican a continuación:

* Mod2Step
* Mod4Setp
* Mod10Back
* Mod12Back

Una vez que hemos elegido los modelos candidatos, como siguiente paso se ha realizado la validación cruzada de los mismos para comprobar su comportamiento y variabilidad con diferentes submuestras de nuestro conjunto de datos y a continuación se incluyen los resultados obtenidos:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Tipo de variables** |  |  | **# variables** |
| Mod2Step | Transformadas | 0.9409202 | 0.0048485 | 33 |
| Mod4Setp | Originales | 0.9384155 | 0.0051780 | 27 |
| Mod10Back | Transformadas | 0.9409917 | 0.0048822 | 35 |
| Mod12Back | Originales | 0.9384567 | 0.0051467 | 29 |

# Como se observa en la tabla anterior los modelos de predicción construidos dan resultados muy similares y la diferencia en la R2 no es muy significativa, por lo que a pesar de que a pesar de que para los modelo Mod2Step y Mod10Back el número de variables utilizadas para la predicción de nuestra variable objetivo es ligeramente superior al resto de modelos, son los modelos que mejor predicen nuestra variable objetivo y su variabilidad es inferior, por lo que estos dos modelos han sido elegidos como candidatos bajo la selección de variables clásica.

# 5.3 Selección de variables aleatoria

Una vez que hemos procedido a la selección de los modelos candidatos, se ha procedido a la construcción de los modelos de regresión lineal a través de la selección de variables aleatoria.

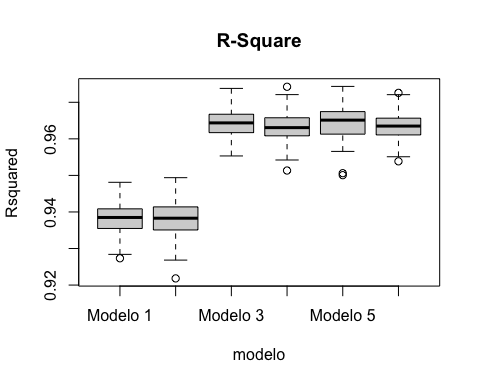
La construcción del modelo de regresión lineal a través de la selección de variables aleatoria se ha realizado considerando 100 repeticiones y una submuestra equivalente al 70% de nuestro conjunto de datos, por lo que los modelos que se han generado mas veces y que han sido elegidos como candidatos para incluirse en la validación cruzada final son los siguientes:

* **Modelo1:**Age\_over65\_pct+CCAA+CCAA:Age\_over65\_pct+CCAA:DifComAutonPtge+CCAA:raiz4Explotaciones+CCAA:SameComAutonPtge+DifComAutonPtge+raiz4Explotaciones+SameComAutonPtge (se ha repetido 4 veces)
* **Modelo2:**Age\_over65\_pct+CCAA+CCAA:Age\_over65\_pct+CCAA:DifComAutonPtge+CCAA:SameComAutonPtge+DifComAutonPtge+raiz4Explotaciones+SameComAutonPtge (se ha repetido 3 veces)
* **Modelo3:**Age\_0.4\_Ptge+Age\_19\_65\_pct+CCAA+CCAA:Age\_19\_65\_pct+CCAA:DifComAutonPtge+CCAA:raiz4Explotaciones+CCAA:SameComAutonPtge+DifComAutonPtge+raiz4Explotaciones+SameComAutonPtge (se ha repetido 2 veces)
* **Modelo4:**Age\_over65\_pct+CCAA+CCAA:Age\_over65\_pct+CCAA:DifComAutonPtge+CCAA:SameComAutonPtge+DifComAutonPtge+raiz4Explotaciones+SameComAutonDiffProvPtge+SameComAutonPtge (se ha repetido 2 veces)

El resto de modelos se han repetido solo una vez por lo que han sido descartados.

Ya que hemos elegido nuestros modelos candidatos, el siguiente paso ha sido realizar la validación cruzada de dichos modelos, incluyendo los dos modelos elegidos como candidatos en la selección de variables clásica con el fin de seleccionar el modelo ganador.

A continuación se incluyen los resultados obtenidos en la validación cruzada para los modelos candidatos bajo ambas metodologías de selección de variables:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Tipo de selección de variables** |  |  | **# variables** |
| Modelo 1 | Clásica | 0.9382263 | 0.004382 | 33 |
| Modelo 2 | Clásica | 0.9383093 | 0.005110 | 35 |
| Modelo 3 | Aleatoria | 0.9643487 | 0.004052 | 85 |
| Modelo 4 | Aleatoria | 0.9632847 | 0.004234 | 69 |
| Modelo 5 | Aleatoria | 0.9643870 | 0.004452 | 86 |
| Modelo 6 | Aleatoria | 0.9633453 | 0.003535 | 70 |

Como se puede observar en el gráfico y en la tabla de resultados, los modelos cuya selección de variables se ha realizado aleatoriamente, son los modelos que mejor predicen nuestra variable objetivo, sin embargo, el número de variables que se utilizan para la predicción es muy elevado.

Por otro lado, aunque los modelos construidos a través de la selección de variables clásica tienen una menor , el número de variables que utiliza estos modelos para predecir nuestra variable objetivo es muy inferior al de el resto de modelos, por lo que se puede concluir que por el principio de parsimonia, el modelo ganador es el Modelo 1 en el cual se incluyen el menor número de variables para su construcción y su variabilidad es mas baja comparada a la del otro modelo construido a través de la selección de variables clásica.

# 5.4 Modelo ganador

Como se ha mencionado en el apartado anterior, el modelo seleccionado como ganador ha sido elegido en función del número de variables incluidas para la predicción de nuestro variable objetivo ya que comparado con otros modelos elegidos como candidatos, la diferencia en su no difiere significativamente del resto de modelos, adicionalmente, se ha demostrado que la variabilidad del modelo no es significativa y por tanto la fiabilidad del mismo es alta.

A continuación se incluye la información del modelo ganador:

## Call:  
## lm(formula = mod1\_clasica, data = data\_train\_t)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53.998 -2.265 -0.117 2.256 48.488   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.350e+01 3.847e+00 -6.107 1.07e-09 \*\*\*  
## CCAAAragón 1.427e+00 4.390e-01 3.249 0.001162 \*\*   
## CCAAAsturias 2.253e+00 8.532e-01 2.641 0.008288 \*\*   
## CCAABaleares 6.990e+00 9.370e-01 7.460 9.78e-14 \*\*\*  
## CCAACanarias 1.293e+01 7.980e-01 16.208 < 2e-16 \*\*\*  
## CCAACantabria 1.409e+00 7.418e-01 1.899 0.057574 .   
## CCAACastillaLeón 5.004e-01 3.535e-01 1.416 0.156960   
## CCAACastillaMancha 2.273e+00 3.980e-01 5.712 1.16e-08 \*\*\*  
## CCAACataluña 7.133e+01 3.763e-01 189.547 < 2e-16 \*\*\*  
## CCAAComValenciana 2.690e+01 4.354e-01 61.790 < 2e-16 \*\*\*  
## CCAAExtremadura 1.183e+00 4.663e-01 2.537 0.011194 \*   
## CCAAGalicia 1.901e+01 4.949e-01 38.411 < 2e-16 \*\*\*  
## CCAAMadrid 2.680e+00 6.299e-01 4.255 2.12e-05 \*\*\*  
## CCAANavarra 1.986e+01 5.250e-01 37.834 < 2e-16 \*\*\*  
## CCAAOtros 2.250e+00 1.021e+00 2.203 0.027609 \*   
## CCAAPaísVasco 5.766e+01 5.428e-01 106.217 < 2e-16 \*\*\*  
## CCAARioja 1.502e+00 6.136e-01 2.448 0.014387 \*   
## SameComAutonPtge 2.371e-01 2.206e-02 10.745 < 2e-16 \*\*\*  
## logxinmuebles 7.907e-01 2.058e-01 3.841 0.000123 \*\*\*  
## Age\_over65\_pct -1.186e-01 1.222e-02 -9.713 < 2e-16 \*\*\*  
## ComercTTEHosteleria1 -4.917e+00 1.252e+00 -3.927 8.71e-05 \*\*\*  
## Age\_0.4\_Ptge 4.380e-01 6.501e-02 6.738 1.75e-11 \*\*\*  
## logxTotalCensus -6.709e+00 9.438e-01 -7.108 1.30e-12 \*\*\*  
## sqrxDifComAutonPtge 5.863e-01 4.228e-02 13.866 < 2e-16 \*\*\*  
## DifComAutonPtge -3.560e-01 3.125e-02 -11.390 < 2e-16 \*\*\*  
## AgricultureUnemploymentPtge -4.241e-02 6.298e-03 -6.734 1.79e-11 \*\*\*  
## expxSameComAutonPtge -1.427e-03 2.697e-04 -5.292 1.25e-07 \*\*\*  
## ActividadPpalOtro -2.549e+00 1.249e+00 -2.042 0.041196 \*   
## ActividadPpalServicios 1.994e+00 3.510e-01 5.682 1.39e-08 \*\*\*  
## DensidadBaja 2.209e+00 3.917e-01 5.639 1.78e-08 \*\*\*  
## DensidadMuyBaja 2.441e+00 4.096e-01 5.958 2.68e-09 \*\*\*  
## UnemployLess25\_Ptge -2.945e-02 8.466e-03 -3.479 0.000507 \*\*\*  
## CodigoProvincia 1.234e-01 2.681e-02 4.604 4.22e-06 \*\*\*  
## sqrxCodigoProvincia -4.521e-01 1.130e-01 -4.002 6.35e-05 \*\*\*  
## logxPob2010 5.484e+00 9.498e-01 5.774 8.10e-09 \*\*\*  
## raiz4PobChange\_pct 6.162e+00 1.701e+00 3.623 0.000294 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.162 on 6460 degrees of freedom  
## Multiple R-squared: 0.94, Adjusted R-squared: 0.9397   
## F-statistic: 2894 on 35 and 6460 DF, p-value: < 2.2e-16

Como se puede observar en lo información del modelo, dentro de la categoría de Comunidad Autónoma, hay categorías que no son significativas en nuestro modelo, por lo que se ha realizado una prueba en donde las CCAA de Cantabria, Castilla y León fueron unidas a la categoría de Otros, sin embargo, la del modelo seleccionado disminuía, por lo tanto, no se ha realizado ningún ajuste adicional a nuestro modelo seleccionado como ganador.

Para la validación del modelo, se han utilizado los datos de entrenamiento y prueba en el que se ha obtenido un del 0.9415632 para los datos de entrenamiento y un de 0.9268544, por lo tanto, podemos concluir que nuestro modelo funciona con distinto conjunto de datos.