Práctica Minería de Datos y Modelización Predictiva

Depuracion de datos

Una ves Depurado todos los los datos los guardamos en un archivo para utilizar esos dato depurados en el modelo final. Necesitamos igualmente ajustar los Datatypes de nuestro FugaClientes_test para que sea compatible con nuestro modelo. Convertimos las siguientes columnas a factor al igual que en nuestro FugaClientes_Training.

Regresion Logistica

Insertamos nuestros datos depurados para entrenar nuestro modelo de regresion logistica.

```
datos<-readRDS("~/Desktop/DS UCM /Modulo 6/Documentación minería de Datos y Modelización
Predictiva-20220129/Tarea/FugaClientes_Training_DEP.RDS")

varObjBin<-datos$varObjBin
input<-datos[,(2:20)]
```

Insertamos una variable aleatoria y veremos si tiene impacto significativo en el modelo. Resulta que la variable aleatoria no aparenta tener impacto suficiente. Las primeras 4 variables significativas son : *contrato*, *Antiguedad*,*Int_serv* y *MetodoPago*

Añadimos nuevas variables transformadas.

Una vez añadido las variables transformadas las primeras cuatro variables en nuestro Vcramer son: *Contrato*,*sqrtxAntiguedad*, *Antiguedad*,*Int_serv*

Buscamos la Frecuencia de 0 y 1 en nuestros datos:

En este caso, tenemos la situación de desbalanceo hacia los 0 ya que la frecuencia a priori de 0 es del 73%. El modelo tendrá mayor dificultad en reconocer a los 1.

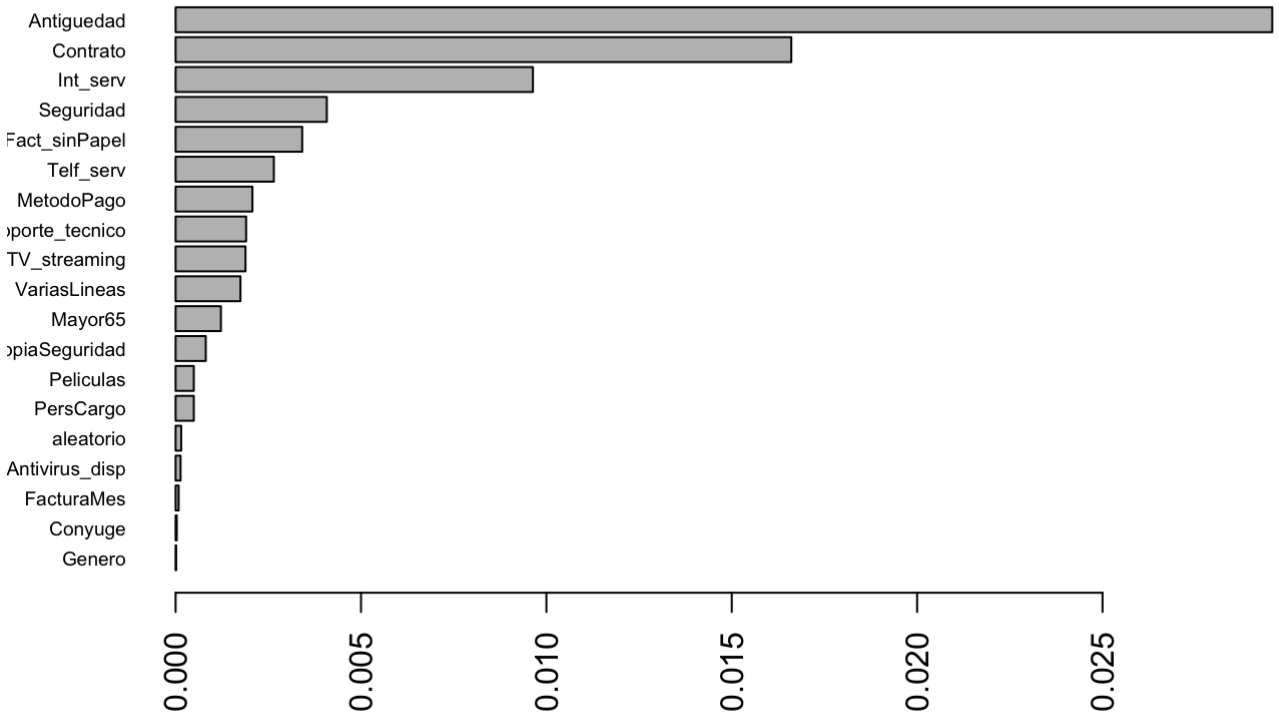
Creamos Modelos

Comenzamos con nuestro modelo completo de referencia que incluye todas las variables.Sin incluir el ID.

En el siguiente apartado vemos las importancia de las variables en orden descendiente. Significando que entre mayor es el Pseudo-R2, mayor su importacia para predecir se van ala Fuga.

```
#pseudoR2(modeloInicial,data_train,"varObjBin")
#pseudoR2(modeloInicial,data_test,"varObjBin")
#modeloInicial$rank #número de parámetros

impVariablesLog(modeloInicial,"varObjBin")
```



Intentamos con los siguientes modelos:

Este modelo es sencillo y bastante significativo en cuanto a sus parámetros. Notamos que el pseudoR2 es mayor en Training versus Test.

```
modelo3<-glm(varObjBin~Antiguedad+Contrato+Int_serv+Seguridad,data=data_train,family=bin
omial)
summary(modelo3)
```

```
##
## Call:
## glm(formula = varObjBin ~ Antiguedad + Contrato + Int_serv +
##     Seguridad, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6016  -0.7117  -0.3220   0.8129   2.9873
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.184510   0.077970  -2.366    0.018 *
## Antiguedad          -0.029305   0.002136 -13.718  < 2e-16 ***
## ContratoOne year    -0.898256   0.116482  -7.712 1.24e-14 ***
## ContratoTwo year    -1.601550   0.173501  -9.231  < 2e-16 ***
## Int_servFiber optic  1.151374   0.086964  13.240  < 2e-16 ***
## Int_servNo          -0.912970   0.130017  -7.022 2.19e-12 ***
## SeguridadYes        -0.613115   0.097872  -6.264 3.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4382.7  on 5076  degrees of freedom
## AIC: 4396.7
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modelo3,data_train,"varObjBin")#No parece muy buena idea
```

```
## [1] 0.2549376
```

```
pseudoR2(modelo3,data_test,"varObjBin")
```

```
## [1] 0.2218762
```

```
modelo3$rank
```

```
## [1] 7
```

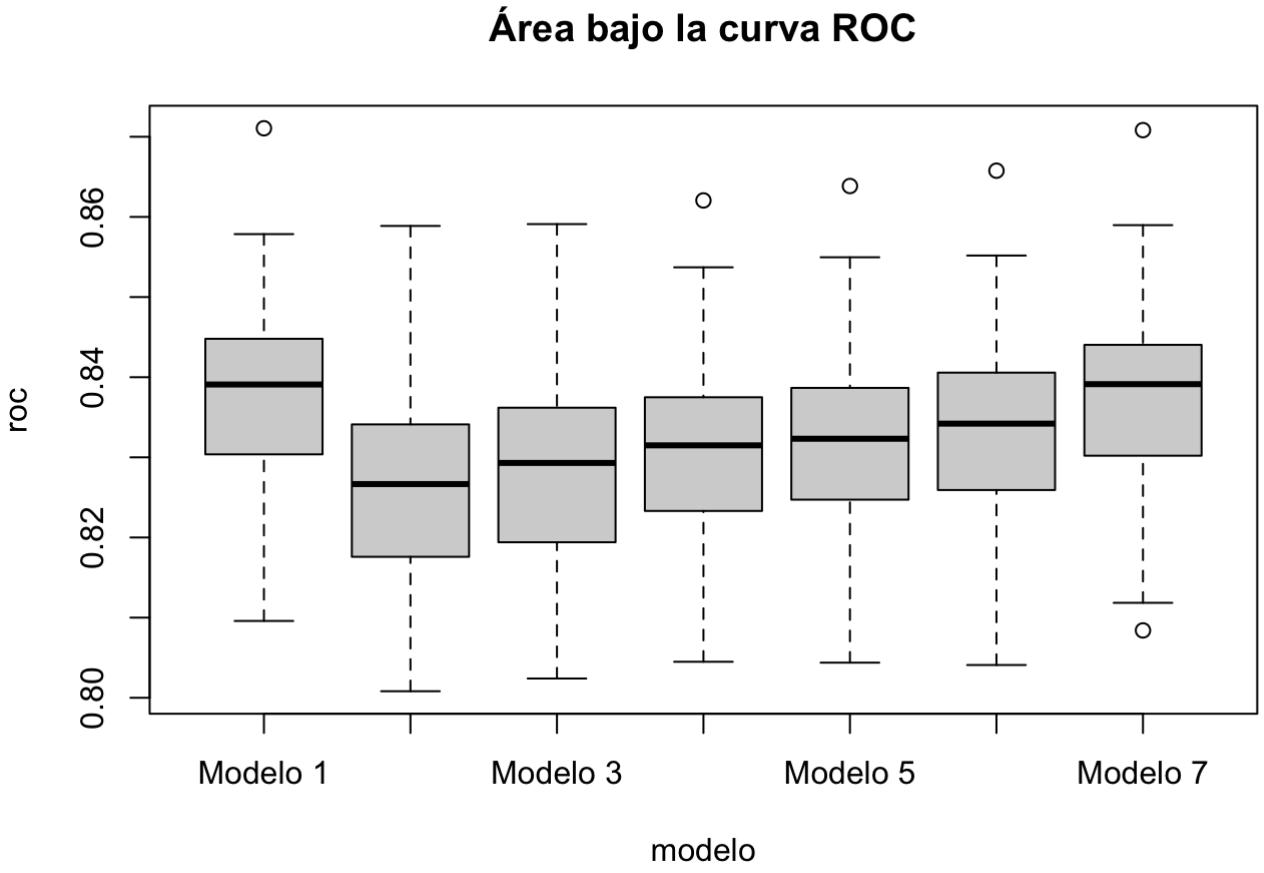
Evaluación de los modelos por validación cruzada repetida

```
#Validacion cruzada repetida para elegir entre todos

#copia de la variable original
auxVarObj<-todo$varObjBin

#formateo la variable objetivo para que funcione el codigo
todo$varObjBin<-make.names(todo$varObjBin)

total<-c()
modelos<-sapply(list(modeloInicial,modelo2,modelo3,modelo4,modelo5,modelo6,modelo7),form
ula)
for (i in 1:length(modelos)){
  set.seed(1712)
  vcr<-train(as.formula(modelos[[i]]), data = todo,
             method = "glm", family="binomial",metric = "ROC",
             trControl = trainControl(method="repeatedcv", number=5, repeats=20,
                                      summaryFunction=twoClassSummary,
                                      classProbs=TRUE,returnResamp="all")
  )
  total<-rbind(total,data.frame(roc=vcr$resample[,1],modelo=rep(paste("Modelo",i),
                                                                nrow(vcr$resample))))
}
boxplot(roc~modelo,data=total,main="Área bajo la curva ROC")
```



```
aggregate(roc~modelo, data = total, mean)
```

modelo <chr>	roc <dbl>
Modelo 1	0.8379688
Modelo 2	0.8259982
Modelo 3	0.8287313
Modelo 4	0.8307887
Modelo 5	0.8318432
Modelo 6	0.8331265
Modelo 7	0.8379545

7 rows

```
aggregate(roc~modelo, data = total, sd)
```

modelo <chr>	roc <dbl>
Modelo 1	0.01172029
Modelo 2	0.01175950
Modelo 3	0.01162390
Modelo 4	0.01164907
Modelo 5	0.01179442
Modelo 6	0.01194238
Modelo 7	0.01176951

7 rows

Para nuestro beneficio hemos seleccionado el modelo3 dado que tiene un valor AIC significativo y un roc de 0.82. Ha pesar de que los otros modelos tengan un valor AIC menor que nuestro modelo3, Hemos decidido este modelo dado su simplicidad con tan solo 7 variables y la diferencia entre pseudoR2_train y pseudoR2_test es tan solo centesimas.

Modelos Mediante seleccion de Variables

Procedemos a la lectura de los datos depurados y con las transformaciones creadas en el código de regresión lineal.

```
# Parto de los datos con las transformaciones creado en el código de regresión lineal
datos<-todo
```

El modelo ganador contenía las variables *Antiguedad*, *Contrato*, *Int_serv* y *Seguridad* y fue el modelo3.

```
# Este fue el modelo manual ganador
modeloManual<- modelo3
summary(modeloManual)
```

```
##
## Call:
## glm(formula = varObjBin ~ Antiguedad + Contrato + Int_serv +
##     Seguridad, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6016  -0.7117  -0.3220   0.8129   2.9873
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.184510   0.077970  -2.366    0.018 *
## Antiguedad          -0.029305   0.002136 -13.718  < 2e-16 ***
## ContratoOne year    -0.898256   0.116482  -7.712 1.24e-14 ***
## ContratoTwo year    -1.601550   0.173501  -9.231  < 2e-16 ***
## Int_servFiber optic  1.151374   0.086964  13.240  < 2e-16 ***
## Int_servNo          -0.912970   0.130017  -7.022 2.19e-12 ***
## SeguridadYes        -0.613115   0.097872  -6.264 3.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4382.7  on 5076  degrees of freedom
## AIC: 4396.7
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloManual,data_train,"varObjBin")
```

```
## [1] 0.2549376
```

```
pseudoR2(modeloManual,data_test,"varObjBin")
```

```
## [1] 0.2218762
```

Selección de variables clásica con variables originales

```
# Seleccion de variables "clásica"
null<-glm(varObjBin~1, data=data_train,family=binomial) #Modelo minimo
full<-glm(varObjBin~., data=data_train,family=binomial) #Modelo maximo

modeloStepAIC<-step(null, scope=list(lower=null, upper=full), direction="both", trace =
F)
summary(modeloStepAIC)
```

```
##
## Call:
## glm(formula = varObjBin ~ Contrato + Int_serv + Antiguedad +
##     Seguridad + Fact_sinPapel + TV_streaming + Soporte_tecnico +
##     Telf_serv + VariasLineas + MetodoPago + Mayor65 + PersCargo +
##     CopiaSeguridad + Peliculas, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0143  -0.6820  -0.3009   0.6849   3.1859
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.095764   0.172017  -0.557 0.577722
## ContratoOne year                -0.819413   0.120082  -6.824 8.87e-12 ***
## ContratoTwo year                -1.442150   0.178062  -8.099 5.53e-16 ***
## Int_servFiber optic              0.951745   0.106284   8.955  < 2e-16 ***
## Int_servNo                      -0.540452   0.149309  -3.620 0.000295 ***
## Antiguedad                      -0.032479   0.002451 -13.250  < 2e-16 ***
## SeguridadYes                    -0.483002   0.100402  -4.811 1.50e-06 ***
## Fact_sinPapelYes                 0.388787   0.086295   4.505 6.63e-06 ***
## TV_streamingYes                  0.352395   0.094764   3.719 0.000200 ***
## Soporte_tecnicoYes              -0.338259   0.101021  -3.348 0.000813 ***
## Telf_servYes                    -0.605997   0.149206  -4.061 4.88e-05 ***
## VariasLineasYes                  0.313408   0.091279   3.434 0.000596 ***
## MetodoPagoCredit card (automatic) -0.143502   0.131499  -1.091 0.275150
## MetodoPagoElectronic check       0.226982   0.109787   2.067 0.038689 *
## MetodoPagoMailed check           0.008627   0.128168   0.067 0.946332
## Mayor651                         0.265090   0.100655   2.634 0.008447 **
## PersCargoYes                    -0.199726   0.094389  -2.116 0.034346 *
## CopiaSeguridadYes               -0.191871   0.090503  -2.120 0.034002 *
## PeliculasYes                     0.188558   0.093483   2.017 0.043692 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4240.0  on 5064  degrees of freedom
## AIC: 4278
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloStepAIC,data_test,"varObjBin")
```

```
## [1] 0.2386129
```

```
modeloBackAIC<-step(full, scope=list(lower=null, upper=full), direction="backward", trac
e = F)
summary(modeloBackAIC)
```

```
##
## Call:
## glm(formula = varObjBin ~ Mayor65 + PersCargo + Antiguedad +
##     Telf_serv + VariasLineas + Int_serv + Seguridad + CopiaSeguridad +
##     Soporte_tecnico + TV_streaming + Peliculas + Contrato + Fact_sinPapel +
##     MetodoPago, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0143  -0.6820  -0.3009   0.6849   3.1859
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.095764   0.172017  -0.557 0.577722
## Mayor651                         0.265090   0.100655   2.634 0.008447 **
## PersCargoYes                    -0.199726   0.094389  -2.116 0.034346 *
## Antiguedad                      -0.032479   0.002451 -13.250  < 2e-16 ***
## Telf_servYes                    -0.605997   0.149206  -4.061 4.88e-05 ***
## VariasLineasYes                  0.313408   0.091279   3.434 0.000596 ***
## Int_servFiber optic              0.951745   0.106284   8.955  < 2e-16 ***
## Int_servNo                      -0.540452   0.149309  -3.620 0.000295 ***
## SeguridadYes                    -0.483002   0.100402  -4.811 1.50e-06 ***
## CopiaSeguridadYes               -0.191871   0.090503  -2.120 0.034002 *
## Soporte_tecnicoYes              -0.338259   0.101021  -3.348 0.000813 ***
## TV_streamingYes                  0.352395   0.094764   3.719 0.000200 ***
## PeliculasYes                     0.188558   0.093483   2.017 0.043692 *
## ContratoOne year                -0.819413   0.120082  -6.824 8.87e-12 ***
## ContratoTwo year                -1.442150   0.178062  -8.099 5.53e-16 ***
## Fact_sinPapelYes                 0.388787   0.086295   4.505 6.63e-06 ***
## MetodoPagoCredit card (automatic) -0.143502   0.131499  -1.091 0.275150
## MetodoPagoElectronic check       0.226982   0.109787   2.067 0.038689 *
## MetodoPagoMailed check           0.008627   0.128168   0.067 0.946332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4240.0  on 5064  degrees of freedom
## AIC: 4278
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloBackAIC,data_test,"varObjBin") #son iguales
```

```
## [1] 0.2386129
```

```
modeloStepBIC<-step(null, scope=list(lower=null, upper=full), direction="both",k=log(nro
w(data_train)), trace = F)
summary(modeloStepBIC)
```

```
##
## Call:
## glm(formula = varObjBin ~ Contrato + Int_serv + Antiguedad +
##     Seguridad + Fact_sinPapel + TV_streaming + Soporte_tecnico +
##     Telf_serv + VariasLineas + Mayor65, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9383  -0.6866  -0.3059   0.7100   3.2133
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.055072   0.139947  -0.394 0.693936
## ContratoOne year    -0.852805   0.119257  -7.151 8.62e-13 ***
## ContratoTwo year    -1.500662   0.176819  -8.487  < 2e-16 ***
## Int_servFiber optic  1.025797   0.104308   9.834  < 2e-16 ***
## Int_servNo          -0.565020   0.145228  -3.891 0.000100 ***
## Antiguedad          -0.034296   0.002341 -14.648  < 2e-16 ***
## SeguridadYes        -0.516583   0.099678  -5.183 2.19e-07 ***
## Fact_sinPapelYes     0.406979   0.085690   4.749 2.04e-06 ***
## TV_streamingYes      0.430090   0.088954   4.835 1.33e-06 ***
## Soporte_tecnicoYes  -0.365369   0.100204  -3.646 0.000266 ***
## Telf_servYes        -0.636404   0.147728  -4.308 1.65e-05 ***
## VariasLineasYes      0.318027   0.090491   3.514 0.000441 ***
## Mayor651             0.326033   0.098665   3.304 0.000952 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4267.6  on 5070  degrees of freedom
## AIC: 4293.6
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloStepBIC,data_test,"varObjBin")
```

```
## [1] 0.2227547
```

```
modeloBackBIC<-step(full, scope=list(lower=null, upper=full), direction="backward",k=log
(nrow(data_train)), trace = F)
summary(modeloBackBIC)
```

```
##
## Call:
## glm(formula = varObjBin ~ Mayor65 + Antiguedad + Telf_serv +
##     VariasLineas + Int_serv + Seguridad + Soporte_tecnico + TV_streaming +
##     Contrato + Fact_sinPapel, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9383  -0.6866  -0.3059   0.7100   3.2133
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.055072   0.139947  -0.394 0.693936
## Mayor651             0.326033   0.098665   3.304 0.000952 ***
## Antiguedad          -0.034296   0.002341 -14.648  < 2e-16 ***
## Telf_servYes        -0.636404   0.147728  -4.308 1.65e-05 ***
## VariasLineasYes      0.318027   0.090491   3.514 0.000441 ***
## Int_servFiber optic  1.025797   0.104308   9.834  < 2e-16 ***
## Int_servNo          -0.565020   0.145228  -3.891 0.000100 ***
## SeguridadYes        -0.516583   0.099678  -5.183 2.19e-07 ***
## Soporte_tecnicoYes  -0.365369   0.100204  -3.646 0.000266 ***
## TV_streamingYes      0.430090   0.088954   4.835 1.33e-06 ***
## ContratoOne year    -0.852805   0.119257  -7.151 8.62e-13 ***
## ContratoTwo year    -1.500662   0.176819  -8.487  < 2e-16 ***
## Fact_sinPapelYes     0.406979   0.085690   4.749 2.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4267.6  on 5070  degrees of freedom
## AIC: 4293.6
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloBackBIC,data_test,"varObjBin") # son iguales
```

```
## [1] 0.2227547
```

```
modeloStepAIC$rank
```

```
## [1] 19
```

```
modeloStepBIC$rank
```

```
## [1] 13
```

Selección de variables clásica con variables originales y transformaciones

```
# Pruebo con todas las transf
fullT<-glm(varObjBin~., data=data_train,family = binomial)

modeloStepAIC_trans<-step(null, scope=list(lower=null, upper=fullT), direction="both", t
race = F)
summary(modeloStepAIC_trans)
```

```
##
## Call:
## glm(formula = varObjBin ~ Contrato + Int_serv + Antiguedad +
##     Seguridad + Fact_sinPapel + TV_streaming + Soporte_tecnico +
##     Telf_serv + VariasLineas + MetodoPago + Mayor65 + PersCargo +
##     CopiaSeguridad + Peliculas, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0143  -0.6820  -0.3009   0.6849   3.1859
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.095764   0.172017  -0.557 0.577722
## ContratoOne year                -0.819413   0.120082  -6.824 8.87e-12 ***
## ContratoTwo year                -1.442150   0.178062  -8.099 5.53e-16 ***
## Int_servFiber optic              0.951745   0.106284   8.955  < 2e-16 ***
## Int_servNo                      -0.540452   0.149309  -3.620 0.000295 ***
## Antiguedad                      -0.032479   0.002451 -13.250  < 2e-16 ***
## SeguridadYes                    -0.483002   0.100402  -4.811 1.50e-06 ***
## Fact_sinPapelYes                 0.388787   0.086295   4.505 6.63e-06 ***
## TV_streamingYes                  0.352395   0.094764   3.719 0.000200 ***
## Soporte_tecnicoYes              -0.338259   0.101021  -3.348 0.000813 ***
## Telf_servYes                    -0.605997   0.149206  -4.061 4.88e-05 ***
## VariasLineasYes                  0.313408   0.091279   3.434 0.000596 ***
## MetodoPagoCredit card (automatic) -0.143502   0.131499  -1.091 0.275150
## MetodoPagoElectronic check       0.226982   0.109787   2.067 0.038689 *
## MetodoPagoMailed check           0.008627   0.128168   0.067 0.946332
## Mayor651                         0.265090   0.100655   2.634 0.008447 **
## PersCargoYes                    -0.199726   0.094389  -2.116 0.034346 *
## CopiaSeguridadYes               -0.191871   0.090503  -2.120 0.034002 *
## PeliculasYes                     0.188558   0.093483   2.017 0.043692 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4240.0  on 5064  degrees of freedom
## AIC: 4278
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloStepAIC_trans,data_test,"varObjBin")
```

```
## [1] 0.2386129
```

```
modeloStepBIC_trans<-step(null, scope=list(lower=null, upper=fullT), direction="both",k=
log(nrow(data_train)), trace = F)
summary(modeloStepBIC_trans)
```

```
##
## Call:
## glm(formula = varObjBin ~ Contrato + Int_serv + Antiguedad +
##     Seguridad + Fact_sinPapel + TV_streaming + Soporte_tecnico +
##     Telf_serv + VariasLineas + Mayor65, family = binomial, data = data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9383  -0.6866  -0.3059   0.7100   3.2133
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.055072   0.139947  -0.394 0.693936
## ContratoOne year    -0.852805   0.119257  -7.151 8.62e-13 ***
## ContratoTwo year    -1.500662   0.176819  -8.487  < 2e-16 ***
## Int_servFiber optic  1.025797   0.104308   9.834  < 2e-16 ***
## Int_servNo          -0.565020   0.145228  -3.891 0.000100 ***
## Antiguedad          -0.034296   0.002341 -14.648  < 2e-16 ***
## SeguridadYes        -0.516583   0.099678  -5.183 2.19e-07 ***
## Fact_sinPapelYes     0.406979   0.085690   4.749 2.04e-06 ***
## TV_streamingYes      0.430090   0.088954   4.835 1.33e-06 ***
## Soporte_tecnicoYes  -0.365369   0.100204  -3.646 0.000266 ***
## Telf_servYes        -0.636404   0.147728  -4.308 1.65e-05 ***
## VariasLineasYes      0.318027   0.090491   3.514 0.000441 ***
## Mayor651             0.326033   0.098665   3.304 0.000952 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5882.3  on 5082  degrees of freedom
## Residual deviance: 4267.6  on 5070  degrees of freedom
## AIC: 4293.6
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoR2(modeloStepBIC_trans,data_test,"varObjBin")
```

```
## [1] 0.2227547
```

```
modeloStepAIC_trans$rank
```

```
## [1] 19
```

```
modeloStepBIC_trans$rank
```

```
## [1] 13
```

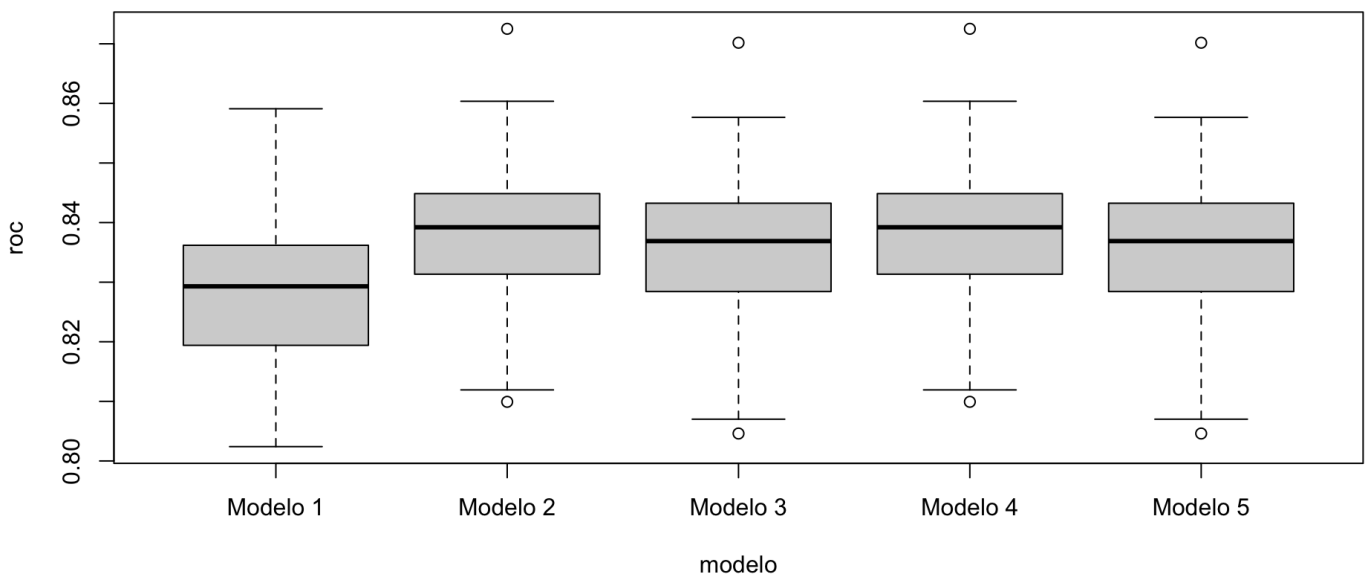
Evaluación por validación cruzada repetida de los modelos se selección clásica

```
#Validacion cruzada repetida para elegir entre todos

#copia de la variable original
auxVarObj<-todo$varObjBin

#formateo la variable objetivo para que funcione el codigo
todo$varObjBin<-make.names(todo$varObjBin)

total<-c()
modelos<-sapply(list(modeloManual,modeloStepAIC,modeloStepBIC,modeloStepAIC_trans,modelo
StepBIC_trans),formula)
for (i in 1:length(modelos)){
  set.seed(1712)
  vcr<-train(as.formula(modelos[[i]]), data = todo,
             method = "glm", family="binomial",metric = "ROC",
             trControl = trainControl(method="repeatedcv", number=5, repeats=20,
                                      summaryFunction=twoClassSummary,
                                      classProbs=TRUE,returnResamp="all")
  )
  total<-rbind(total,data.frame(roc=vcr$resample[,1],modelo=rep(paste("Modelo",i),
                                                                nrow(vcr$resample))))
}
boxplot(roc~modelo,data=total,main="Área bajo la curva ROC")
```




```
aggregate(roc~modelo, data = total, mean)
```

modelo <chr>	roc <dbl>
Modelo 1	0.8287313
Modelo 2	0.8385406
Modelo 3	0.8356890
Modelo 4	0.8385406
Modelo 5	0.8356890

5 rows

```
aggregate(roc~modelo, data = total, sd)
```

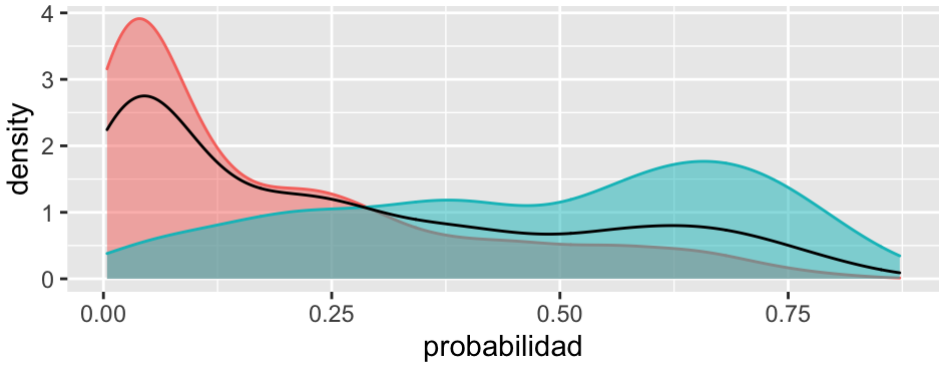
modelo <chr>	roc <dbl>
Modelo 1	0.01162390
Modelo 2	0.01176733
Modelo 3	0.01201366
Modelo 4	0.01176733
Modelo 5	0.01201366

5 rows

En base a los nuevos modelos creados en el apartado de crear modelos mediante 'seleccion de variables', Hemos decidido selccionar nuestro modeloStepAIC cuyo tiene mejor AIC (4284.7) y demuestra tener mejora area bajo la curva con un roc de 0.893. Entendemos que este modelo tendra mejor capacidad predictiva. lo cual comprobamos a continuacion.

Punto de corte óptimo para la probabilidad estimada

```
#gráfico de las probabilidades obtenidas
hist_targetbinaria(predict(modeloStepAIC, newdata=data_test,type="response"),data_test$v
arObjBin,"probabilidad")
```



Observamos nuestra distribucionde probabilidad y densidad de valores 0(rojo) y 1(azul).

Segun la interpretacion de nuestro modelo: determinamos que los siguientes puntos de corte para modeloStepAIC

```
## generamos una rejilla de puntos de corte
posiblesCortes<-seq(0,1,0.01)
rejilla<-data.frame(t(rbind(posiblesCortes,sapply(posiblesCortes,function(x) sensEspCort
e(modeloStepAIC,data_test,"varObjBin",x,"1")))))
rejilla$Youden<-rejilla$Sensitivity+rejilla$Specificity-1
#plot(rejilla$posiblesCortes,rejilla$Youden)
#plot(rejilla$posiblesCortes,rejilla$Accuracy)
rejilla$posiblesCortes[which.max(rejilla$Youden)]
```

```
## [1] 0.3
```

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

```
## [1] 0.54
```

```
#Comprobamos los puntos de corte
sensEspCorte(modeloStepAIC,data_test,"varObjBin",0.33,"1")
```

```
##       Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value
##      0.7669291      0.7091988      0.7877814      0.5469108      0.8823529
```

```
sensEspCorte(modeloStepAIC,data_test,"varObjBin",0.53,"1")
```

```
##       Accuracy    Sensitivity    Specificity Pos Pred Value Neg Pred Value
##      0.7921260      0.4836795      0.9035370      0.6442688      0.8289086
```

Vemos que nos dan puntos de corte maximo 0.33 y 0.53. El corte 0.33 nos da un Sensitivity(probabilidad de detectar los 1) de 0.706 balanceada con la Specificity(probabilida de detectar los 0) de 0.788.A diferencia del punto de corte de 0.52 que queda desbalanceado la Specificity= 0.913 versus Sensitivity.= 0.48. Por el cual elegimos 0.33 como nuestro puntode corte.

#Prueba con Martiz

Comprobamos nuestro modelo:

Probamos Nuestro Modelo elegido contra data_test. Como notamos que nuestra data_test tenia realmente 337 Fugas. y nuestro modelo ha detectado 435. Una diferencia de sobre-estimacion de 116 fugas. Como nuestro obetivo es prevenir las Fugas de los clienes es mejor sobre-estimar los valores suponiendo que el costo de perder al cliente es mayor que el costo en prevenir su fuga. Nuesto modelo de prueba tiene un *sensitivity* de 70% de detectar las fugas mientras un 78% para dectectaar los No-Fuga, con un Accuracy total de 77% en el modelo y un Kappa de 45% el cual indica que nuestro modelo de regresion es competente.

TEST

Ahora ponemos a trabajar a nuestro modelo seleccionado con nuestras predicciones. Recordemos que el 1=FUGA y 0=NO_FUGA

```
pred_test<-factor(ifelse(predict(modeloStepAIC,FugaClientes_test,type = "response")>0.33
,1,0))

# Tablas marginales
table(pred_test)
```

```
## pred_test
##   0   1
## 448 242
```

Para la data de Fuga_test nuestro modelo ha detectado una proporcionde 64% de los clientes No-Fuga y 36% Fugan. Una proporcion similar a la data que utliziamos de prueba cuya proporcion era mayor para los No-Fuga.

Convertimos nuestros resultados a un Dataset y expotamos la data a un RDS.