Análisis del dataset Diabetes

Sherly Tarazona

28/6/2021

### Objetivos:

* Clasificar el reingreso de un paciente con diabetes
* Identificar patrones que permitan anticipar un reingreso

##### Nota: Los resultados obtenidos deberian ser revisados en conjunto con médicos especialistas

#Libraries  
library(MASS)  
library(nnet)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(scorecard)  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(ggplot2)  
library(knitr)  
library(reshape2)  
library(skimr)

## Read dataset

data<-read.csv("diabetic\_data.csv", header=T, sep=",")  
df<-data

## Exploring dataset

#duplicados  
nrow(df) - nrow(unique(df))

## [1] 0

#Type of features:  
#Permite observar de forma general cada una de las variables: missing, valores unicos, minimos, maximos, etc.  
skim(df)

Data summary

|  |  |
| --- | --- |
| Name | df |
| Number of rows | 101766 |
| Number of columns | 50 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 37 |
| numeric | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| race | 0 | 1 | 1 | 15 | 0 | 6 | 0 |
| gender | 0 | 1 | 4 | 15 | 0 | 3 | 0 |
| age | 0 | 1 | 6 | 8 | 0 | 10 | 0 |
| weight | 0 | 1 | 1 | 9 | 0 | 10 | 0 |
| payer\_code | 0 | 1 | 1 | 2 | 0 | 18 | 0 |
| medical\_specialty | 0 | 1 | 1 | 36 | 0 | 73 | 0 |
| diag\_1 | 0 | 1 | 1 | 6 | 0 | 717 | 0 |
| diag\_2 | 0 | 1 | 1 | 6 | 0 | 749 | 0 |
| diag\_3 | 0 | 1 | 1 | 6 | 0 | 790 | 0 |
| max\_glu\_serum | 0 | 1 | 4 | 4 | 0 | 4 | 0 |
| A1Cresult | 0 | 1 | 2 | 4 | 0 | 4 | 0 |
| metformin | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| repaglinide | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| nateglinide | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| chlorpropamide | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| glimepiride | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| acetohexamide | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| glipizide | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| glyburide | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| tolbutamide | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| pioglitazone | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| rosiglitazone | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| acarbose | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| miglitol | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| troglitazone | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| tolazamide | 0 | 1 | 2 | 6 | 0 | 3 | 0 |
| examide | 0 | 1 | 2 | 2 | 0 | 1 | 0 |
| citoglipton | 0 | 1 | 2 | 2 | 0 | 1 | 0 |
| insulin | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| glyburide.metformin | 0 | 1 | 2 | 6 | 0 | 4 | 0 |
| glipizide.metformin | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| glimepiride.pioglitazone | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| metformin.rosiglitazone | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| metformin.pioglitazone | 0 | 1 | 2 | 6 | 0 | 2 | 0 |
| change | 0 | 1 | 2 | 2 | 0 | 2 | 0 |
| diabetesMed | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| readmitted | 0 | 1 | 2 | 3 | 0 | 3 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| encounter\_id | 0 | 1 | 165201645.62 | 102640295.98 | 12522 | 84961194 | 152388987 | 230270888 | 443867222 | ▆▇▅▂▂ |
| patient\_nbr | 0 | 1 | 54330400.69 | 38696359.35 | 135 | 23413221 | 45505143 | 87545950 | 189502619 | ▇▆▆▁▁ |
| admission\_type\_id | 0 | 1 | 2.02 | 1.45 | 1 | 1 | 1 | 3 | 8 | ▇▂▁▁▁ |
| discharge\_disposition\_id | 0 | 1 | 3.72 | 5.28 | 1 | 1 | 1 | 4 | 28 | ▇▁▁▁▁ |
| admission\_source\_id | 0 | 1 | 5.75 | 4.06 | 1 | 1 | 7 | 7 | 25 | ▅▇▁▁▁ |
| time\_in\_hospital | 0 | 1 | 4.40 | 2.99 | 1 | 2 | 4 | 6 | 14 | ▇▅▂▁▁ |
| num\_lab\_procedures | 0 | 1 | 43.10 | 19.67 | 1 | 31 | 44 | 57 | 132 | ▃▇▅▁▁ |
| num\_procedures | 0 | 1 | 1.34 | 1.71 | 0 | 0 | 1 | 2 | 6 | ▇▂▁▁▁ |
| num\_medications | 0 | 1 | 16.02 | 8.13 | 1 | 10 | 15 | 20 | 81 | ▇▃▁▁▁ |
| number\_outpatient | 0 | 1 | 0.37 | 1.27 | 0 | 0 | 0 | 0 | 42 | ▇▁▁▁▁ |
| number\_emergency | 0 | 1 | 0.20 | 0.93 | 0 | 0 | 0 | 0 | 76 | ▇▁▁▁▁ |
| number\_inpatient | 0 | 1 | 0.64 | 1.26 | 0 | 0 | 0 | 1 | 21 | ▇▁▁▁▁ |
| number\_diagnoses | 0 | 1 | 7.42 | 1.93 | 1 | 6 | 8 | 9 | 16 | ▁▅▇▁▁ |

##Data cleaning

#Eliminando variables que no aportan  
df$encounter\_id<-NULL  
df$patient\_nbr<-NULL  
  
#Data cleaning  
df$race[df$race=="?"]<-"None"  
df$weight[df$weight=="?"]<- "None"  
df$payer\_code[df$payer\_code=="?"]<-"None"  
df$medical\_specialty[df$medical\_specialty=="?"]<-"None"  
  
df$admission\_type\_id<-as.character(df$admission\_type\_id)  
df$admission\_source\_id<-as.character(df$admission\_source\_id)

#### Defining Target

#Para este analisis, consideraremos como categoria objetivo a aquellos que han reingresado desde 1 dia en adelante  
df$target\_y<-ifelse(df$readmitted=="NO",0,1)  
table(df$target\_y)

##   
## 0 1   
## 54864 46902

### Sampling

#Testing & training set  
  
dt\_list = split\_df(df, y="target\_y", ratio = c(0.7, 0.3), seed = 3000)  
test<-dt\_list$test  
train<-dt\_list$train  
dim(train)

## [1] 71114 49

dim(test)

## [1] 30652 49

round(prop.table(table(train$target\_y))\*100,2)

##   
## 0 1   
## 54 46

#Selection:  
  
features<-c(  
 "target\_y"  
 ,"race"  
 ,"gender"  
 ,"age"  
 ,"weight"  
 ,"payer\_code"  
 #,"medical\_specialty"  
 ,"max\_glu\_serum"  
 ,"A1Cresult"  
 ,"metformin"  
 ,"repaglinide"  
 ,"nateglinide"  
 ,"chlorpropamide"  
 ,"glimepiride"  
 ,"acetohexamide"  
 ,"glipizide"  
 ,"glyburide"  
 ,"tolbutamide"  
 ,"pioglitazone"  
 ,"rosiglitazone"  
 ,"acarbose"  
 ,"miglitol"  
 ,"troglitazone"  
 ,"tolazamide"  
 ,"insulin"  
 ,"glyburide.metformin"  
 ,"glipizide.metformin"  
 ,"glimepiride.pioglitazone"  
 ,"metformin.rosiglitazone"  
 ,"metformin.pioglitazone"  
 ,"change"  
 ,"diabetesMed"  
   
#numeric   
 ,"admission\_type\_id"  
 ,"discharge\_disposition\_id"  
 ,"admission\_source\_id"  
 ,"time\_in\_hospital"  
 ,"num\_lab\_procedures"  
 ,"num\_procedures"  
 ,"num\_medications"  
 ,"number\_outpatient"  
 ,"number\_emergency"  
 ,"number\_inpatient"  
 ,"number\_diagnoses"   
)

### Binning

train<-as.data.frame(train)  
test<-as.data.frame(test)  
  
  
#Bineado  
bins <- woebin(train[,features], y = "target\_y",   
 #min\_perc\_fine\_bin=0.02, # How many bns to cut initially into  
 min\_perc\_coarse\_bin=0.05, # Minimum percentage per final bin  
 stop\_limit=0.02, # Minimum information value   
 max\_num\_bin=6, method='tree')

## [INFO] creating woe binning ...

## Warning in check\_const\_cols(dt): There were 1 constant columns removed from input dataset,  
## metformin.pioglitazone

ncol(train[,features])

## [1] 42

### Feature selection

iv <- vector("numeric", 0)  
iv\_names <- vector("numeric", 0)  
#val=seq(1,ncol(train[,features])-1)  
val=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40)  
  
for(i in val) {  
 iv[i]=round(as.data.frame(bins[i])[,c(1,10)][1,2],3)  
 iv\_names[i]=as.data.frame(bins[i])[,c(1,10)][1,1]  
}  
a<-as.data.frame(cbind(iv\_names,iv))  
knitr::kable(a[order(a$iv,decreasing=T),])

|  | iv\_names | iv |
| --- | --- | --- |
| 39 | number\_inpatient | 0.224 |
| 38 | number\_emergency | 0.063 |
| 40 | number\_diagnoses | 0.056 |
| 37 | number\_outpatient | 0.049 |
| 32 | admission\_source\_id | 0.045 |
| 31 | discharge\_disposition\_id | 0.037 |
| 36 | num\_medications | 0.033 |
| 5 | payer\_code | 0.018 |
| 23 | insulin | 0.018 |
| 33 | time\_in\_hospital | 0.018 |
| 29 | diabetesMed | 0.016 |
| 30 | admission\_type\_id | 0.014 |
| 3 | age | 0.01 |
| 1 | race | 0.009 |
| 35 | num\_procedures | 0.009 |
| 28 | change | 0.008 |
| 34 | num\_lab\_procedures | 0.008 |
| 8 | metformin | 0.003 |
| 7 | A1Cresult | 0.002 |
| 14 | glipizide | 0.002 |
| 2 | gender | 0.001 |
| 18 | rosiglitazone | 0.001 |
| 4 | weight | 0 |
| 6 | max\_glu\_serum | 0 |
| 9 | repaglinide | 0 |
| 10 | nateglinide | 0 |
| 11 | chlorpropamide | 0 |
| 12 | glimepiride | 0 |
| 13 | acetohexamide | 0 |
| 15 | glyburide | 0 |
| 16 | tolbutamide | 0 |
| 17 | pioglitazone | 0 |
| 19 | acarbose | 0 |
| 20 | miglitol | 0 |
| 21 | troglitazone | 0 |
| 22 | tolazamide | 0 |
| 24 | glyburide.metformin | 0 |
| 25 | glipizide.metformin | 0 |
| 26 | glimepiride.pioglitazone | 0 |
| 27 | metformin.rosiglitazone | 0 |

### Modeling: Logit

train\_woe = woebin\_ply(train[,features], bins)

## [INFO] converting into woe values ...

test\_woe = woebin\_ply(test[,features], bins)

## [INFO] converting into woe values ...

#Target como factor  
train\_woe$target\_y<-as.factor(train\_woe$target\_y)  
  
#Modelos  
model1 = glm( train\_woe$target\_y ~   
   
 number\_inpatient\_woe  
 +number\_emergency\_woe  
 +number\_diagnoses\_woe  
 +number\_outpatient\_woe  
 #+medical\_specialty\_woe  
 +admission\_source\_id\_woe  
 +discharge\_disposition\_id\_woe  
 +num\_medications\_woe  
 +payer\_code\_woe  
 +insulin\_woe  
 +time\_in\_hospital\_woe  
 +diabetesMed\_woe  
 +admission\_type\_id\_woe  
 +age\_woe  
   
 #poco aporte:  
 #+race\_woe  
 #+num\_procedures\_woe  
 #+change\_woe  
 #+num\_lab\_procedures\_woe  
 #+metformin\_woe  
 #+A1Cresult\_woe  
 #+glipizide\_woe  
 #+gender\_woe  
 #+rosiglitazone\_woe  
   
 ,family = binomial(link="logit"), data = train\_woe)  
  
knitr::kable(vif(model1, merge\_coef = TRUE), format="simple")

| variable | Estimate | Std. Error | z value | Pr(>|z|) | gvif |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | -0.1584895 | 0.0079 | -20.1319 | 0.0000 | NA |
| number\_inpatient\_woe | 0.8343298 | 0.0175 | 47.7713 | 0.0000 | 1.057811 |
| number\_emergency\_woe | 0.4729504 | 0.0330 | 14.3103 | 0.0000 | 1.067831 |
| number\_diagnoses\_woe | 0.3874455 | 0.0362 | 10.6993 | 0.0000 | 1.169833 |
| number\_outpatient\_woe | 0.5037129 | 0.0367 | 13.7400 | 0.0000 | 1.048044 |
| admission\_source\_id\_woe | 0.6394115 | 0.0415 | 15.4129 | 0.0000 | 1.214226 |
| discharge\_disposition\_id\_woe | 0.7042191 | 0.0413 | 17.0647 | 0.0000 | 1.026066 |
| num\_medications\_woe | 0.3510564 | 0.0476 | 7.3813 | 0.0000 | 1.210267 |
| payer\_code\_woe | 0.4485071 | 0.0609 | 7.3634 | 0.0000 | 1.055725 |
| insulin\_woe | 0.1852532 | 0.0645 | 2.8713 | 0.0041 | 1.210450 |
| time\_in\_hospital\_woe | 0.2824194 | 0.0634 | 4.4521 | 0.0000 | 1.160891 |
| diabetesMed\_woe | 0.6870740 | 0.0676 | 10.1604 | 0.0000 | 1.192348 |
| admission\_type\_id\_woe | 0.5142250 | 0.0727 | 7.0769 | 0.0000 | 1.172501 |
| age\_woe | 0.6158581 | 0.0836 | 7.3636 | 0.0000 | 1.086485 |
| ### Feature contribution |  |  |  |  |  |

#Contribucion por variable  
a<-vif(model1, merge\_coef = TRUE)[,c("variable","z value")]  
b<-a[order(a$`z value`,decreasing=T),]  
n=nrow(b)  
total=sum(b[1:(n-1),]$`z value`)  
contribution<-b[1:(n-1),][,2]\*100/total  
knitr::kable(cbind(b[1:(n-1),],"%contribution"=contribution))

| variable | z value | %contribution.z value |
| --- | --- | --- |
| number\_inpatient\_woe | 47.7713 | 28.835650 |
| discharge\_disposition\_id\_woe | 17.0647 | 10.300572 |
| admission\_source\_id\_woe | 15.4129 | 9.303515 |
| number\_emergency\_woe | 14.3103 | 8.637965 |
| number\_outpatient\_woe | 13.7400 | 8.293721 |
| number\_diagnoses\_woe | 10.6993 | 6.458297 |
| diabetesMed\_woe | 10.1604 | 6.133007 |
| num\_medications\_woe | 7.3813 | 4.455491 |
| age\_woe | 7.3636 | 4.444807 |
| payer\_code\_woe | 7.3634 | 4.444686 |
| admission\_type\_id\_woe | 7.0769 | 4.271749 |
| time\_in\_hospital\_woe | 4.4521 | 2.687371 |
| insulin\_woe | 2.8713 | 1.733170 |

### Performance: Auc

auc\_gini<-function(muestra,modelo){  
 pred<-predict(modelo,muestra,type = "response")  
 auc=round(roc(muestra$target\_y,pred)$auc,3)  
 gini=round(2\*auc-1,2)   
 cbind(auc,gini)  
}  
  
train\_<-auc\_gini(muestra=train\_woe,modelo=model1)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

test\_<-auc\_gini(muestra=test\_woe,modelo=model1)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

c<-as.data.frame(rbind(train\_,test\_))  
rownames(c)<-c("train\_","test\_")#,"back\_")  
knitr::kable(c)

|  | auc | gini |
| --- | --- | --- |
| train\_ | 0.667 | 0.33 |
| test\_ | 0.667 | 0.33 |

#Main features

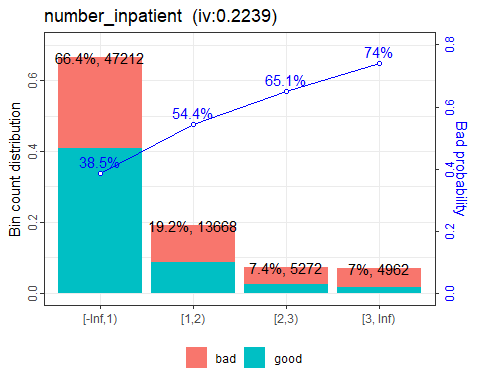
g1<-woebin\_plot(bins$number\_inpatient)  
g2<-woebin\_plot(bins$discharge\_disposition\_id)  
g3<-woebin\_plot(bins$admission\_source\_id)  
#g4<-woebin\_plot(bins$medical\_specialty)  
g5<-woebin\_plot(bins$admission\_source\_id)  
g6<-woebin\_plot(bins$number\_emergency)  
g7<-woebin\_plot(bins$number\_outpatient)  
g8<-woebin\_plot(bins$diabetesMed)  
g9<-woebin\_plot(bins$number\_diagnoses)  
g10<-woebin\_plot(bins$num\_medications)  
#g11<-woebin\_plot(bins$payer\_code)  
g12<-woebin\_plot(bins$age)  
g13<-woebin\_plot(bins$time\_in\_hospital)  
g14<-woebin\_plot(bins$insulin)  
g15<-woebin\_plot(bins$admission\_type\_id)

### Causes

#### Estos tipo de graficos permiten evaluar las categorias con mayor tasa de ocurrencia, a fin de generar alertas para una posible anticipacion al reingreso

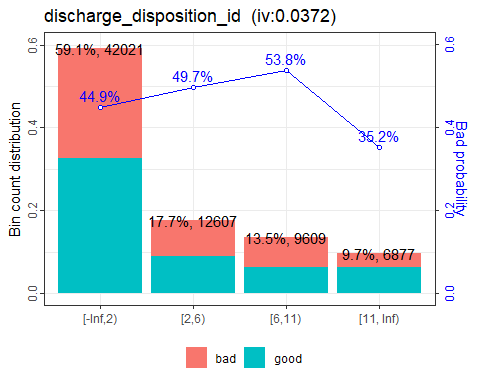
#número de pacientes hospitalizados  
g1

## $number\_inpatient



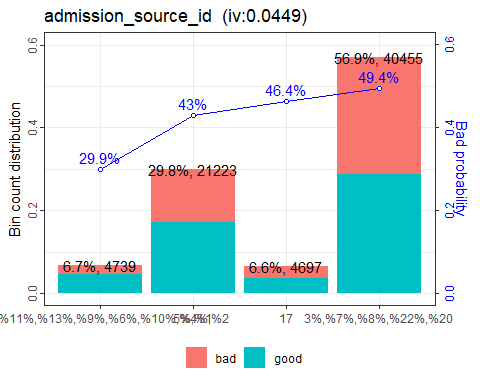
#discharge\_disposition\_id  
#  
g2

## $discharge\_disposition\_id



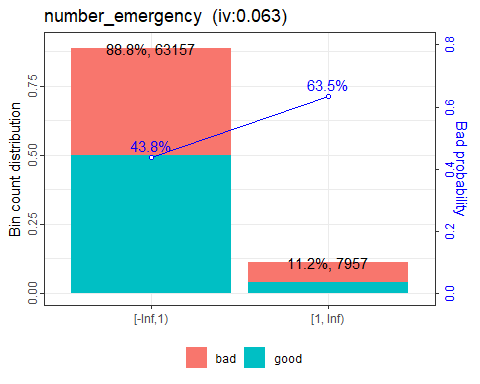
# Motivo de admision  
g5

## $admission\_source\_id



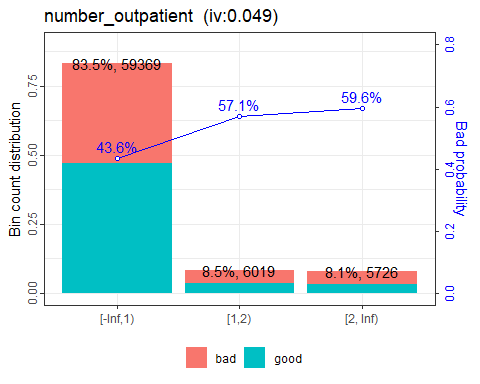
#Numero de pacientes en emergencia  
g6

## $number\_emergency



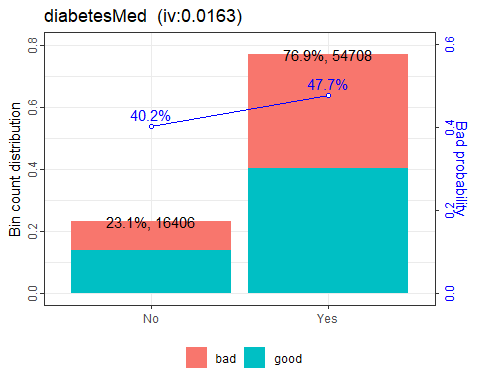
#Numero de pacientes ambulatorios  
g7

## $number\_outpatient



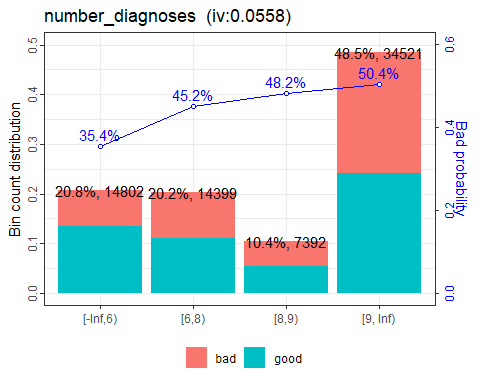
#Estado de Diabeted Med  
g8

## $diabetesMed



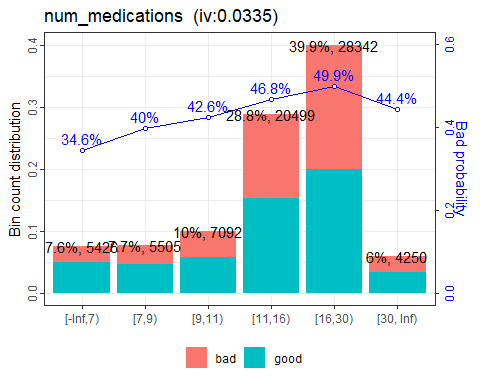
#Numero de diagnosticos  
g9

## $number\_diagnoses



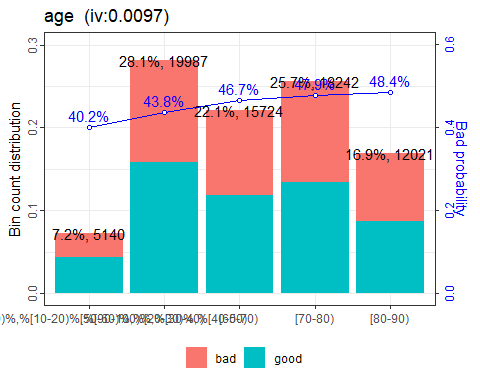
#Numero de medicamentos  
g10

## $num\_medications



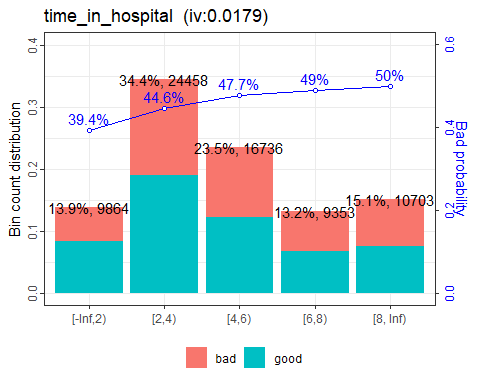
#Edad  
g12

## $age



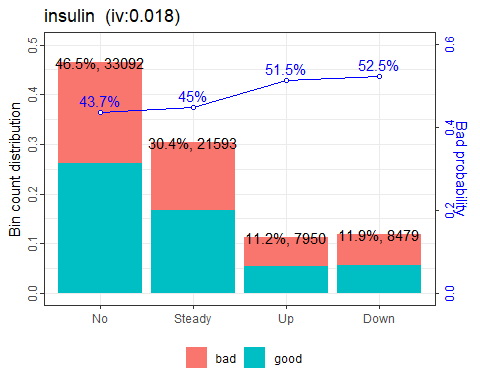
#Tiempo en el hospital  
g13

## $time\_in\_hospital



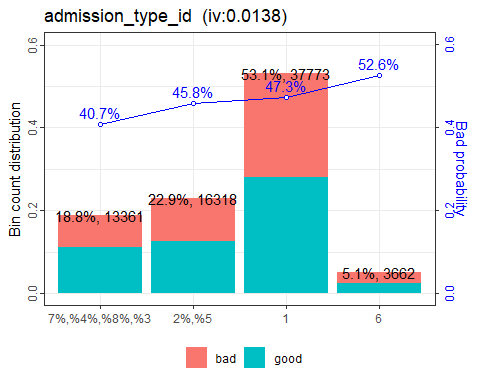
#Estado de la insulina  
g14

## $insulin



# Tipo de admision  
g15

## $admission\_type\_id



### Model 2

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

modelo\_2 <- randomForest(target\_y~.,data=train[,features],importance = TRUE, maxnodes=10,mtry=6,ntree=70)

## Warning in randomForest.default(m, y, ...): The response has five or fewer  
## unique values. Are you sure you want to do regression?

modelo\_2

##   
## Call:  
## randomForest(formula = target\_y ~ ., data = train[, features], importance = TRUE, maxnodes = 10, mtry = 6, ntree = 70)   
## Type of random forest: regression  
## Number of trees: 70  
## No. of variables tried at each split: 6  
##   
## Mean of squared residuals: 0.2345077  
## % Var explained: 5.59

train\_\_<-auc\_gini(muestra=train,modelo=modelo\_2)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

test\_\_<-auc\_gini(muestra=test,modelo=modelo\_2)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

c\_<-as.data.frame(rbind(train\_\_,test\_\_))  
rownames(c\_)<-c("train\_\_","test\_\_")  
knitr::kable(c\_)

|  | auc | gini |
| --- | --- | --- |
| train\_\_ | 0.668 | 0.34 |
| test\_\_ | 0.667 | 0.33 |

#Feature importance  
  
importancia=data.frame(importance(modelo\_2))  
library(reshape)

##   
## Attaching package: 'reshape'

## The following objects are masked from 'package:reshape2':  
##   
## colsplit, melt, recast

## The following object is masked from 'package:dplyr':  
##   
## rename

importancia<-sort\_df(importancia,vars='IncNodePurity')  
varImpPlot(modelo\_2,n.var = 20)

