Churn Prediction

Author: 135430

Introducción

En este trabajo del modulo 5, se va a tratar el caso de Churn Prediction para el dataset de una compañia de telecomunicaciones. Se va a tratar de por un lado entender porque los clientes abandonan la compañia y por otro como detectar esa fuga en la medida de lo posible mediante analisis de los datos y aplicación de algoritmos supervisados.

Repositorio

Todo el código como el dataset se encuentran el el siguiente repositorio de git junto con este mismo notebook en .rmd y html: https://github.com/alexbr86/telco_churn (https://github.com/alexbr86/telco_churn) .

Inicialización

Funciones auxiliares para cargar paquetes y lista de paquetes

```
prepare_packages <- function(packages){
  # Chequeamos que paquetes no estan instalados:
  non_intalled <- packages[!(packages %in% installed.packages()[, "Package"])]
  # En caso de existir alguno aún no instalado, lo instalamos:
  if (length(non_intalled))
    install.packages(non_intalled, dependencies = TRUE)
  # Cargamos toda la lista de paquetes:
  sapply(packages, require, character.only = TRUE)
}
packages <- c("tidyverse",
              "MASS",
              "car",
              "binr",
              "e1071",
              "caret",
              "cowplot",
              "caTools",
              "pROC",
              "ggcorrplot",
              "data.table",
              "Information",
              "rpart",
              "rpart.plot",
              "xgboost",
              "ROCR",
              "pROC",
              "GGally",
              "fastDummies"
)
prepare_packages(packages)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------
------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------
------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: binr
```

```
## Loading required package: e1071
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
## Loading required package: cowplot
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

```
## Loading required package: caTools
```

```
## Loading required package: pROC
```

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

```
##
## Attaching package: 'pROC'
```

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

```
## Loading required package: ggcorrplot
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## Loading required package: Information
```

```
## Loading required package: rpart
```

```
## Loading required package: rpart.plot
```

```
## Loading required package: xgboost
```

```
##
## Attaching package: 'xgboost'
```

```
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
## Loading required package: ROCR
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
## Loading required package: fastDummies
```

```
##   tidyverse        MASS         car        binr       e1071       caret
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##     cowplot     caTools        pROC  ggcorrplot  data.table Information
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##       rpart  rpart.plot     xgboost        ROCR        pROC      GGally
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
## fastDummies
##        TRUE
```

Cargar el dataset

```
dataset <- read.csv("dataset/WA_Fn-UseC_-Telco-Customer-Churn.csv")
```

Se va a analizar el tipo de información que contiene el dataset

```
glimpse(dataset)
```

```
## Observations: 7,043
## Variables: 21
## $ customerID       <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOC...
## $ gender           <fct> Female, Male, Male, Male, Female, Female, Mal...
## $ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Partner          <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes...
## $ Dependents       <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes...
## $ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 5...
## $ PhoneService     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes...
## $ MultipleLines    <fct> No phone service, No, No, No phone service, N...
## $ InternetService  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic,...
## $ OnlineSecurity   <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Yes, ...
## $ OnlineBackup     <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, N...
## $ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, N...
## $ TechSupport      <fct> No, No, No, Yes, No, No, No, No, Yes, No, No,...
## $ StreamingTV      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No...
## $ StreamingMovies  <fct> No, No, No, No, No, Yes, No, No, Yes, No, No,...
## $ Contract         <fct> Month-to-month, One year, Month-to-month, One...
## $ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No,...
## $ PaymentMethod    <fct> Electronic check, Mailed check, Mailed check,...
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89....
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820....
## $ Churn            <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, N...
```

Limpieza del dataset

Se va a cambiar los valores de la variable Senior Citizens valores categoricos de ‘Yes’, ‘No’. Homogeneizando con el resto de variables categoricas.

```
dataset$SeniorCitizen <- as.factor(ifelse(dataset$SeniorCitizen==1, 'Yes', 'No'))
```

A continuación se va a ver un resumen del dataset para poder entenderlo mejor y detectar información necesaria a primera vista.

```
summary(dataset)
```

```
##      customerID      gender     SeniorCitizen Partner    Dependents
##  0002-ORFBO:   1   Female:3488   No :5901      No :3641   No :4933
##  0003-MKNFE:   1   Male  :3555   Yes:1142      Yes:3402   Yes:2110
##  0004-TLHLJ:   1
##  0011-IGKFF:   1
##  0013-EXCHZ:   1
##  0013-MHZWF:   1
##  (Other)   :7037
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##
##              OnlineSecurity              OnlineBackup
##  No                 :3498   No                 :3088
##  No internet service:1526   No internet service:1526
##  Yes                :2019   Yes                :2429
##
##
##
##
##             DeviceProtection              TechSupport
##  No                 :3095    No                 :3473
##  No internet service:1526    No internet service:1526
##  Yes                :2422    Yes                :2044
##
##
##
##
##              StreamingTV              StreamingMovies
##  No                 :2810   No                 :2785
##  No internet service:1526   No internet service:1526
##  Yes                :2707   Yes                :2732
##
##
##
##
##           Contract    PaperlessBilling                   PaymentMethod
##  Month-to-month:3875   No :2872         Bank transfer (automatic):1544
##  One year      :1473   Yes:4171         Credit card (automatic)  :1522
##  Two year      :1695                    Electronic check         :2365
##                                         Mailed check             :1612
##
##
##
##  MonthlyCharges    TotalCharges    Churn
##  Min.   : 18.25   Min.   :  18.8   No :5174
##  1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
##  Median : 70.35   Median :1397.5
##  Mean   : 64.76   Mean   :2283.3
##  3rd Qu.: 89.85   3rd Qu.:3794.7
##  Max.   :118.75   Max.   :8684.8
##                   NA's   :11
```

Tras ver el resumen, puede observarse que la variable TotalCharges tiene 11 NAs que habría que limipiar.Antes de ver que solución se le imputa al problema, es conveniente analizar el porqué son NAs, si es un fallo o hay algún motivo en los datos.

```
dataset[is.na(dataset$TotalCharges),]
```

```
##      customerID gender SeniorCitizen Partner Dependents tenure
## 489  4472-LVYGI Female            No     Yes        Yes      0
## 754  3115-CZMZD   Male            No      No        Yes      0
## 937  5709-LVOEQ Female            No     Yes        Yes      0
## 1083 4367-NUYAO   Male            No     Yes        Yes      0
## 1341 1371-DWPAZ Female            No     Yes        Yes      0
## 3332 7644-OMVMY   Male            No     Yes        Yes      0
## 3827 3213-VVOLG   Male            No     Yes        Yes      0
## 4381 2520-SGTTA Female            No     Yes        Yes      0
## 5219 2923-ARZLG   Male            No     Yes        Yes      0
## 6671 4075-WKNIU Female            No     Yes        Yes      0
## 6755 2775-SEFEE   Male            No      No        Yes      0
##      PhoneService    MultipleLines InternetService      OnlineSecurity
## 489            No No phone service             DSL                 Yes
## 754           Yes               No              No No internet service
## 937           Yes               No             DSL                 Yes
## 1083          Yes              Yes              No No internet service
## 1341           No No phone service             DSL                 Yes
## 3332          Yes               No              No No internet service
## 3827          Yes              Yes              No No internet service
## 4381          Yes               No              No No internet service
## 5219          Yes               No              No No internet service
## 6671          Yes              Yes             DSL                  No
## 6755          Yes              Yes             DSL                 Yes
##             OnlineBackup    DeviceProtection         TechSupport
## 489                   No                 Yes                 Yes
## 754  No internet service No internet service No internet service
## 937                  Yes                 Yes                  No
## 1083 No internet service No internet service No internet service
## 1341                 Yes                 Yes                 Yes
## 3332 No internet service No internet service No internet service
## 3827 No internet service No internet service No internet service
## 4381 No internet service No internet service No internet service
## 5219 No internet service No internet service No internet service
## 6671                 Yes                 Yes                 Yes
## 6755                 Yes                  No                 Yes
##              StreamingTV     StreamingMovies Contract PaperlessBilling
## 489                  Yes                  No Two year              Yes
## 754  No internet service No internet service Two year               No
## 937                  Yes                 Yes Two year               No
## 1083 No internet service No internet service Two year               No
## 1341                 Yes                  No Two year               No
## 3332 No internet service No internet service Two year               No
## 3827 No internet service No internet service Two year               No
## 4381 No internet service No internet service Two year               No
## 5219 No internet service No internet service One year              Yes
## 6671                 Yes                  No Two year               No
## 6755                  No                  No Two year              Yes
##                  PaymentMethod MonthlyCharges TotalCharges Churn
## 489  Bank transfer (automatic)          52.55           NA    No
## 754               Mailed check          20.25           NA    No
## 937               Mailed check          80.85           NA    No
## 1083              Mailed check          25.75           NA    No
## 1341   Credit card (automatic)          56.05           NA    No
## 3332              Mailed check          19.85           NA    No
## 3827              Mailed check          25.35           NA    No
## 4381              Mailed check          20.00           NA    No
```

```
## 5219              Mailed check          19.70           NA    No
## 6671              Mailed check          73.35           NA    No
## 6755 Bank transfer (automatic)          61.90           NA    No
```

Tras analizar los 11 casos de NAs se ha observado que esto es debido a que son clientes recientes y no llevan aún ni un cargo acumulado, ya que como se ve en la columna Tenure llevan 0 meses en la compañia. Por lo que la mejor solución a este problema sería imputarles a todos un valor de 0. Primero se va a comprobar si hay más clientes que lleven 0 meses en la compañia y si tengan valor en TotalCharges

```
dataset[dataset$tenure==0,]
```

```
##      customerID gender SeniorCitizen Partner Dependents tenure
## 489  4472-LVYGI Female            No     Yes        Yes      0
## 754  3115-CZMZD   Male            No      No        Yes      0
## 937  5709-LVOEQ Female            No     Yes        Yes      0
## 1083 4367-NUYAO   Male            No     Yes        Yes      0
## 1341 1371-DWPAZ Female            No     Yes        Yes      0
## 3332 7644-OMVMY   Male            No     Yes        Yes      0
## 3827 3213-VVOLG   Male            No     Yes        Yes      0
## 4381 2520-SGTTA Female            No     Yes        Yes      0
## 5219 2923-ARZLG   Male            No     Yes        Yes      0
## 6671 4075-WKNIU Female            No     Yes        Yes      0
## 6755 2775-SEFEE   Male            No      No        Yes      0
##      PhoneService    MultipleLines InternetService      OnlineSecurity
## 489            No No phone service             DSL                 Yes
## 754           Yes               No              No No internet service
## 937           Yes               No             DSL                 Yes
## 1083          Yes              Yes              No No internet service
## 1341           No No phone service             DSL                 Yes
## 3332          Yes               No              No No internet service
## 3827          Yes              Yes              No No internet service
## 4381          Yes               No              No No internet service
## 5219          Yes               No              No No internet service
## 6671          Yes              Yes             DSL                  No
## 6755          Yes              Yes             DSL                 Yes
##             OnlineBackup    DeviceProtection         TechSupport
## 489                   No                 Yes                 Yes
## 754  No internet service No internet service No internet service
## 937                  Yes                 Yes                  No
## 1083 No internet service No internet service No internet service
## 1341                 Yes                 Yes                 Yes
## 3332 No internet service No internet service No internet service
## 3827 No internet service No internet service No internet service
## 4381 No internet service No internet service No internet service
## 5219 No internet service No internet service No internet service
## 6671                 Yes                 Yes                 Yes
## 6755                 Yes                  No                 Yes
##              StreamingTV     StreamingMovies Contract PaperlessBilling
## 489                  Yes                  No Two year              Yes
## 754  No internet service No internet service Two year               No
## 937                  Yes                 Yes Two year               No
## 1083 No internet service No internet service Two year               No
## 1341                 Yes                  No Two year               No
## 3332 No internet service No internet service Two year               No
## 3827 No internet service No internet service Two year               No
## 4381 No internet service No internet service Two year               No
## 5219 No internet service No internet service One year              Yes
## 6671                 Yes                  No Two year               No
## 6755                  No                  No Two year              Yes
##                  PaymentMethod MonthlyCharges TotalCharges Churn
## 489  Bank transfer (automatic)          52.55           NA    No
## 754               Mailed check          20.25           NA    No
## 937               Mailed check          80.85           NA    No
## 1083              Mailed check          25.75           NA    No
## 1341   Credit card (automatic)          56.05           NA    No
## 3332              Mailed check          19.85           NA    No
## 3827              Mailed check          25.35           NA    No
## 4381              Mailed check          20.00           NA    No
```

```
## 5219              Mailed check          19.70           NA    No
## 6671              Mailed check          73.35           NA    No
## 6755 Bank transfer (automatic)          61.90           NA    No
```

Se comprueba que todos aquellos que llevan 0 meses son solo aquellos que tienen valores NA en TotalCharges. Por lo que se les va a asignar o como valor de pago acumulado en total.

```
#Asignar valor 0 a todos aquellos que sean NA en TotalCharges
dataset[dataset$tenure==0, "TotalCharges"]=0
var_class <- sapply(dataset, class)
var_class_numeric <- names(dataset[var_class=="numeric"])
var_class_inter <- names(dataset[var_class=="integer"])
var_num_total <- c(var_class_numeric, var_class_inter)
var_numeric <- dataset[var_num_total]
```

AutoML

Utilizando la libreria de H2O se va a probar como se comporta el dataset en limpio con un modelo y asi poder intuir de que modo proceder con el mismo.

Se va a dividir el dataset en train, validation y test.

```
set.seed(46)
selected <- sample(1:nrow(dataset), 0.2*nrow(dataset))
train <- dataset[-selected,]
test <- dataset[selected,]
#Model
# Set names for h2o
target <- "Churn"
x <- setdiff(names(train), target)
```

Se va a lanzar todos los modelos supervisados excepto los referentes a Deep Learning y GLM.

```
library(h2o)
```

```
##
## ----------------------------------------------------------------------
##
## Your next step is to start H2O:
##     > h2o.init()
##
## For H2O package documentation, ask for help:
##     > ??h2o
##
## After starting H2O, you can use the Web UI at http://localhost:54321
## For more information visit http://docs.h2o.ai
##
## ----------------------------------------------------------------------
```

```
##
## Attaching package: 'h2o'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, month, week, year
```

```
## The following object is masked from 'package:pROC':
##
##     var
```

```
## The following objects are masked from 'package:stats':
##
##     cor, sd, var
```

```
## The following objects are masked from 'package:base':
##
##     %*%, %in%, &&, ||, apply, as.factor, as.numeric, colnames,
##     colnames<-, ifelse, is.character, is.factor, is.numeric, log,
##     log10, log1p, log2, round, signif, trunc
```

```
h2o.init()
```

```
##  Connection successful!
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         1 hours 9 minutes
##     H2O cluster timezone:       Europe/Paris
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.22.1.1
##     H2O cluster version age:    6 months and 2 days !!!
##     H2O cluster name:           H2O_started_from_R_Alex_zkv489
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   3.19 GB
##     H2O cluster total cores:    8
##     H2O cluster allowed cores:  8
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     H2O API Extensions:         Algos, AutoML, Core V3, Core V4
##     R Version:                  R version 3.6.0 (2019-04-26)
```

```
## Warning in h2o.clusterInfo():
## Your H2O cluster version is too old (6 months and 2 days)!
## Please download and install the latest version from http://h2o.ai/download/
```

```
write.csv(train, file = "train.csv")
train2 = h2o.importFile("./train.csv")
```

```
##
  |
  |                                                                      |   0%
  |
  |=======================================================               |  84%
  |
  |======================================================================| 100%
```

```
write.csv(test, file = "test.csv")
test2 = h2o.importFile("./test.csv")
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

```
aml <- h2o.automl(x = x,
                  y = target,
                  validation_frame = test2,
                  training_frame = train2,
                  max_runtime_secs = 60,
                  exclude_algos = c("DeepLearning", "GLM", "DRF", "StackedEnsemble"))
```

```
##
  |
  |                                                                      |   0%
  |
  |===                                                                   |   5%
  |
  |======                                                                |   9%
  |
  |=========                                                             |  15%
  |
  |============                                                          |  18%
  |
  |================                                                      |  23%
  |
  |==================                                                    |  27%
  |
  |====================                                                  |  31%
  |
  |========================                                              |  35%
  |
  |=========================                                             |  36%
  |
  |============================                                          |  43%
  |
  |==============================                                        |  45%
  |
  |==============================                                        |  46%
  |
  |================================                                      |  47%
  |
  |================================                                      |  49%
  |
  |==================================                                    |  50%
  |
  |====================================                                  |  52%
  |
  |====================================                                  |  54%
  |
  |======================================                                |  55%
  |
  |========================================                              |  56%
  |
  |========================================                              |  57%
  |
  |=========================================                             |  58%
  |
  |=========================================                             |  59%
  |
  |===========================================                           |  61%
  |
  |===========================================                           |  62%
  |
  |============================================                          |  63%
  |
  |============================================                          |  64%
  |
  |=============================================                         |  65%
```

```
|
|=========================================                      |  66%
|
|==========================================                     |  67%
|
|==========================================                     |  68%
|
|===========================================                    |  69%
|
|============================================                   |  71%
|
|============================================                   |  72%
|
|============================================                   |  73%
|
|=============================================                  |  74%
|
|=============================================                  |  75%
|
|==============================================                 |  76%
|
|==============================================                 |  77%
|
|===============================================                |  78%
|
|================================================               |  80%
|
|================================================               |  81%
|
|=================================================              |  82%
|
|=================================================              |  83%
|
|==================================================             |  85%
|
|==================================================             |  86%
|
|===================================================            |  87%
|
|====================================================           |  88%
|
|====================================================           |  90%
|
|=====================================================          |  91%
|
|======================================================         |  92%
|
|======================================================         |  93%
|
|=======================================================        |  94%
|
|========================================================       |  95%
|
|========================================================       |  96%
|
|=========================================================      |  97%
|
|========================================================== |  98%
```

```
  |
  |======================================================================| 100%
```

Extraer los mejores modelos del train.

```
automl_leader <- aml@leader
automl_leader_list <- aml@leaderboard
automl_leader_list
```

```
##                                      model_id       auc   logloss
## 1 GBM_grid_1_AutoML_20190630_224714_model_14 0.8449336 0.4156784
## 2  GBM_grid_1_AutoML_20190630_224714_model_6 0.8432442 0.4397276
## 3               GBM_5_AutoML_20190630_224714 0.8431277 0.4174469
## 4  GBM_grid_1_AutoML_20190630_224714_model_3 0.8423618 0.4469264
## 5 GBM_grid_1_AutoML_20190630_224714_model_11 0.8419958 0.4383928
## 6 GBM_grid_1_AutoML_20190630_224714_model_10 0.8412579 0.4440543
##   mean_per_class_error      rmse       mse
## 1            0.2367116 0.3677093 0.1352101
## 2            0.2365354 0.3760384 0.1414049
## 3            0.2367679 0.3684129 0.1357281
## 4            0.2373333 0.3792955 0.1438651
## 5            0.2330777 0.3754945 0.1409961
## 6            0.2357053 0.3778462 0.1427678
##
## [19 rows x 6 columns]
```

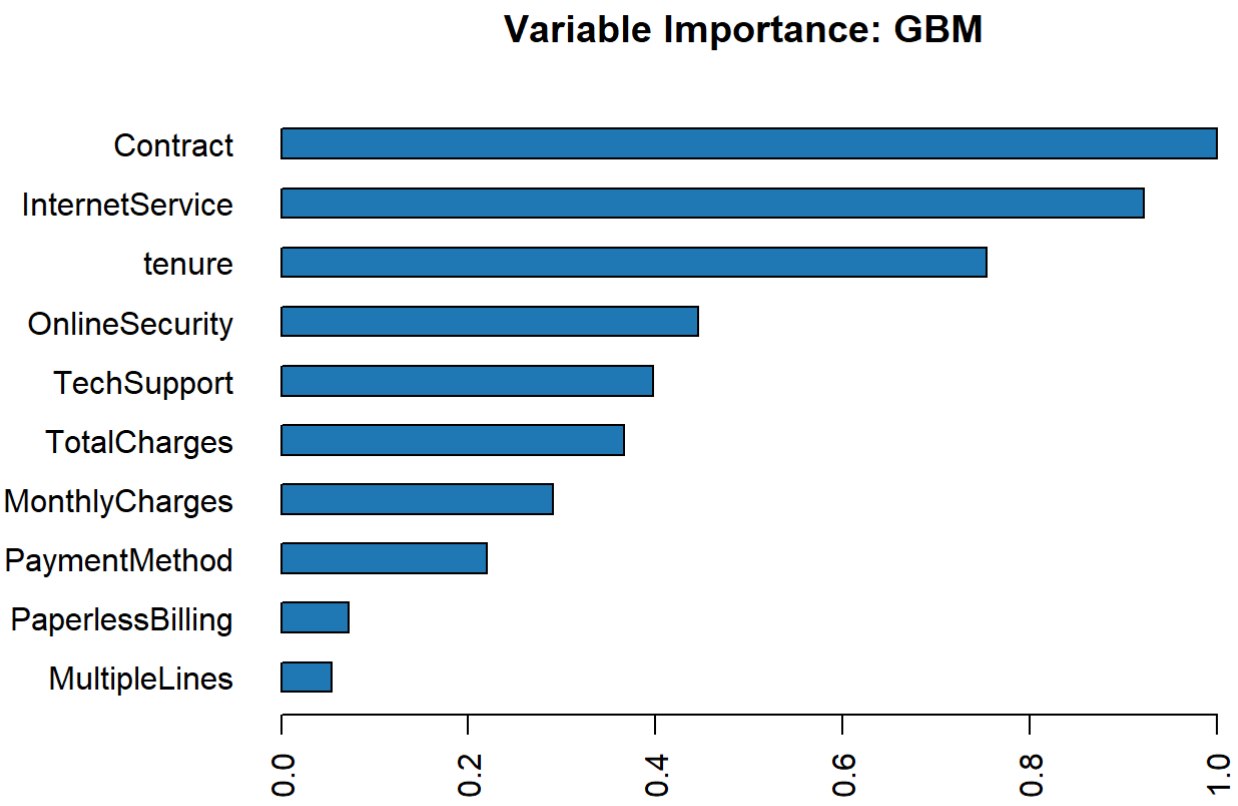
Ver la matriz de confusión

```
h2o.confusionMatrix(automl_leader)
```

```
## Confusion Matrix (vertical: actual; across: predicted)  for max f1 @ threshold = 0.3214113
0065597:
##          No  Yes    Error        Rate
## No     3315  838 0.201782   =838/4153
## Yes     328 1154 0.221323   =328/1482
## Totals 3643 1992 0.206921  =1166/5635
```

Variables más determinantes

```
h2o.varimp_plot(automl_leader)
```



Las variables más determinantes teniendo en cuenta el mejor algoritmo seleccionado por AutoML son: Contract: como variable más determinantes sería la duración del contrato. TechSupport: la segunda más determinante si el cliente tiene soporte técnico. Tenure: Es la tercera más determinante, cuanto tiempo lleva siendo cliente de la compañia.

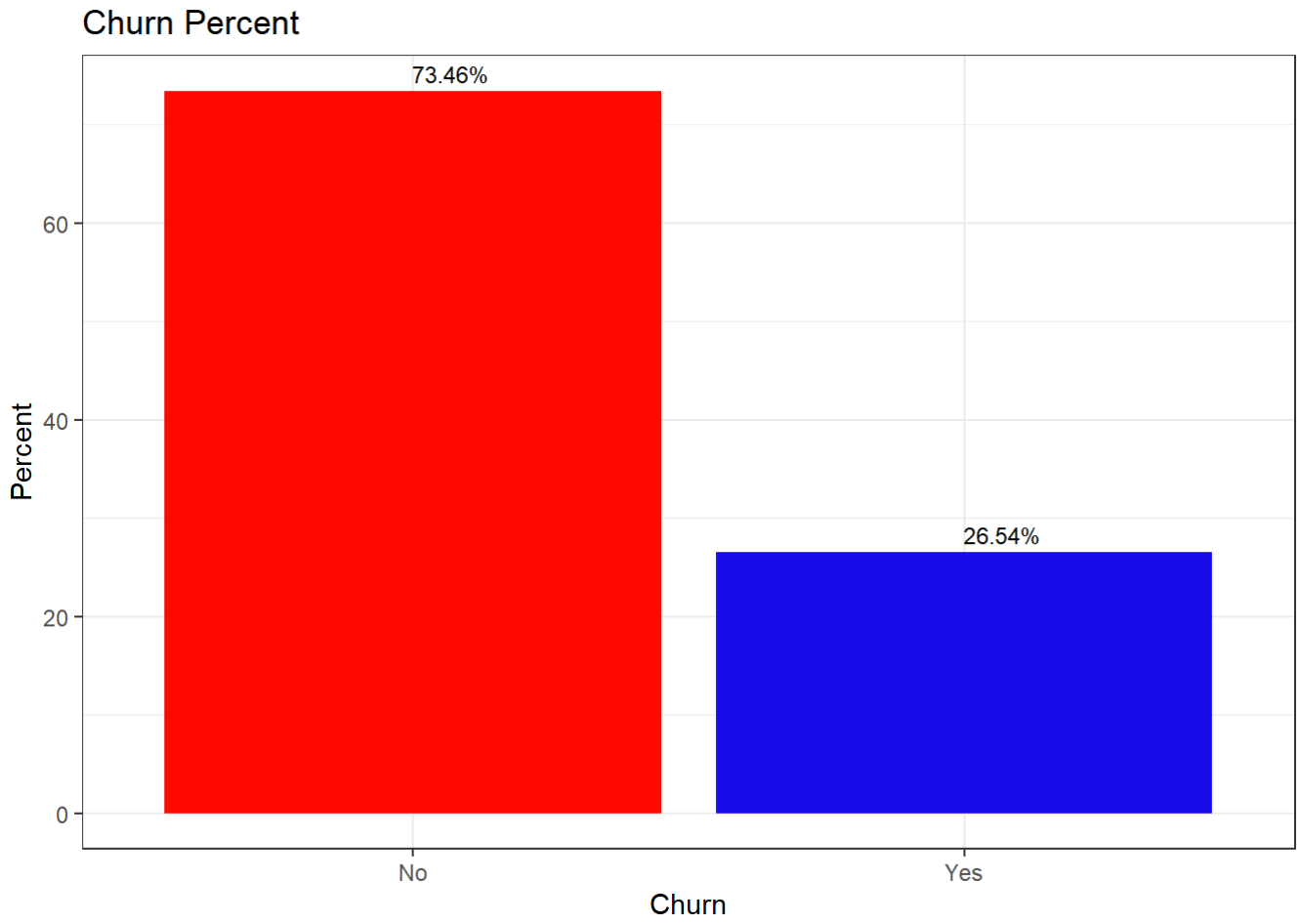
EDA

Ahora se va a analizar el dataset y sus variables. Con ello tratar de describir el dataset y obtener la mayor cantidad de información relevante para luego poder montar el modelo de predicción de abandono.

Variable target Churn

Primero se va a ver que porcentaje del dataset da positivo en abandono:

```
options(repr.plot.width = 6, repr.plot.height = 4)
dataset %>%
group_by(Churn) %>%
summarise(Count = n())%>%
mutate(percent = prop.table(Count)*100)%>%
ggplot(aes(reorder(Churn, -percent), percent), fill = Churn)+
geom_col(fill = c("#FF0800", "#170CEA"))+
geom_text(aes(label = sprintf("%.2f%%", percent)), hjust = 0.01,vjust = -0.5, size =3)+
theme_bw()+
xlab("Churn") +
ylab("Percent")+
ggtitle("Churn Percent")
```

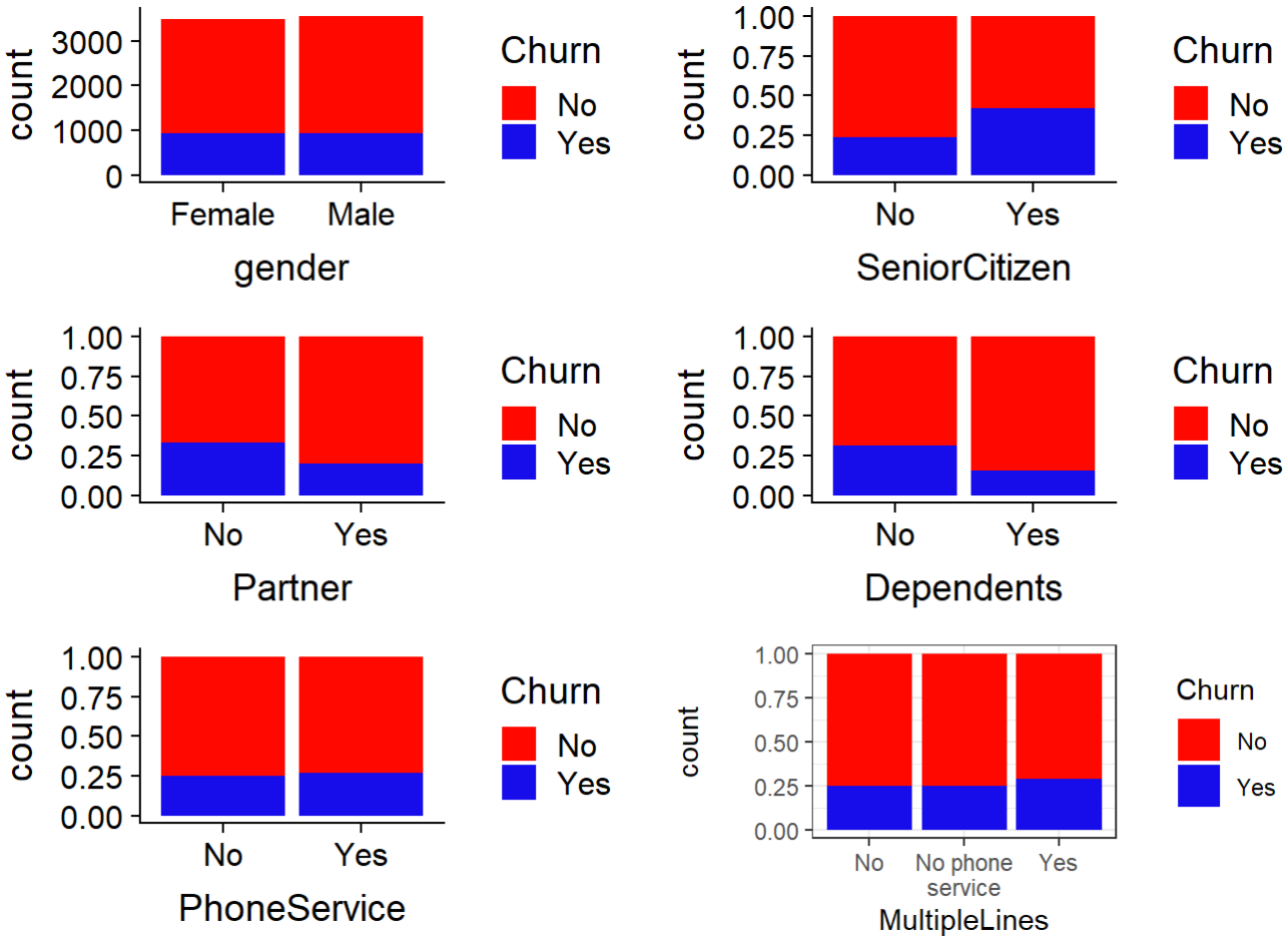


Se puede observar que del dataset algo más de un 25% da positivo en abandono. Es una muestra bastante bien balanceada para lo que suele ser este tipo de casos, por lo que inicialmente podria ser viable descartar hacer down o up sampling.

Variables categoricas

Ahora vamos a ver como se distribuye la variable de Churn en el resto de variables categoricas:

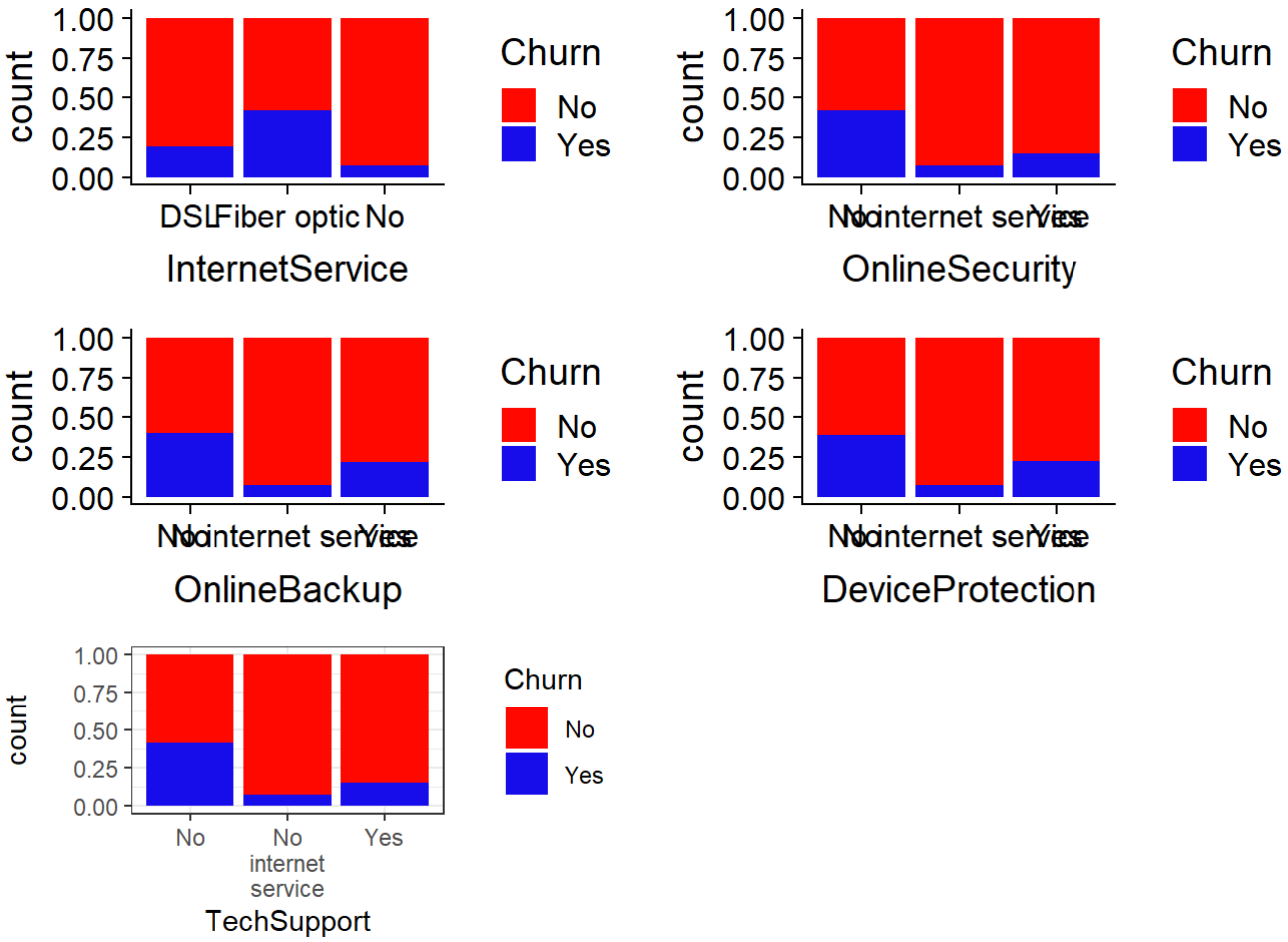
```
options(repr.plot.width = 12, repr.plot.height = 100)
plot_grid(ggplot(dataset, aes(x=gender,fill=Churn)) + geom_bar() + scale_fill_manual(values=c
("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=SeniorCitizen,fill=Churn))+ geom_bar(position = 'fill') + sca
le_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=Partner,fill=Churn))+ geom_bar(position = 'fill') + scale_fil
l_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=Dependents,fill=Churn))+ geom_bar(position = 'fill') + scale_
fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=PhoneService,fill=Churn))+ geom_bar(position = 'fill') + scal
e_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=MultipleLines,fill=Churn))+ geom_bar(position = 'fill') + sca
le_fill_manual(values=c("#FF0800", "#170CEA"))
            + theme_bw()+
          scale_x_discrete(labels = function(x) str_wrap(x, width = 10)),
          align = 'v', ncol=2)
```



Como puede observarse en la mayoria de estas primeras variables se observa un equilibrio entre sus valores con respecto a la variable objetivo Churn. Aunque hay algunas que ya nos dan algun indicador de tendencia como:

- SeniorCitizen, indica que si el cliente es senior o no, y en la gráfica se ve que los clientes senior tienen mayor tendencia a abandonar la compañia.
- Partner, si el cliente tiene un partner o no, y hay una tendencia mayor al abandono entre los que no lo tienen.
- Dependents, si el cliente dependientes, y se observa una tendencia mayor al abandono en los clientes que no tienen dependientes.

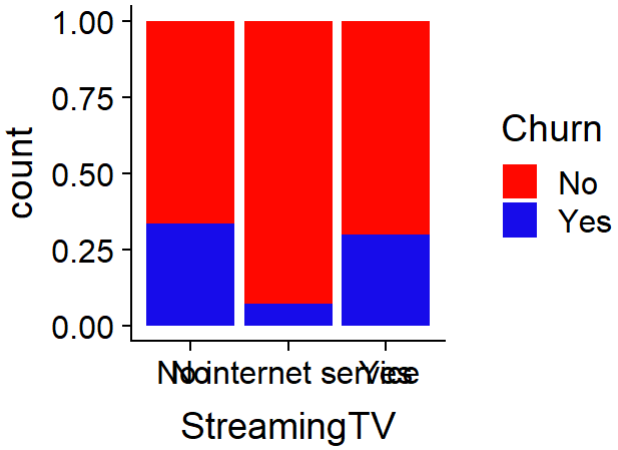
```
options(repr.plot.width = 12, repr.plot.height = 100)
plot_grid(ggplot(dataset, aes(x=InternetService,fill=Churn))+ geom_bar(position = 'fill') + s
cale_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=OnlineSecurity,fill=Churn))+ geom_bar(position = 'fill') + sc
ale_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=OnlineBackup,fill=Churn))+ geom_bar(position = 'fill') + scal
e_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=DeviceProtection,fill=Churn))+ geom_bar(position = 'fill') +
 scale_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=TechSupport,fill=Churn))+ geom_bar(position = 'fill') + scale
_fill_manual(values=c("#FF0800", "#170CEA"))
            + theme_bw()+
          scale_x_discrete(labels = function(x) str_wrap(x, width = 10)),
          align = 'v', ncol=2)
```

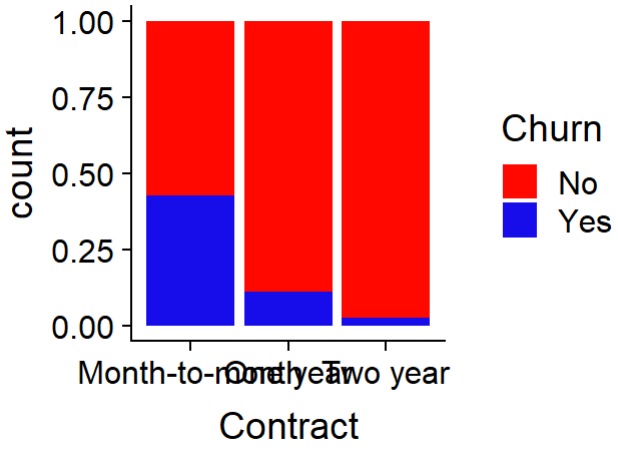


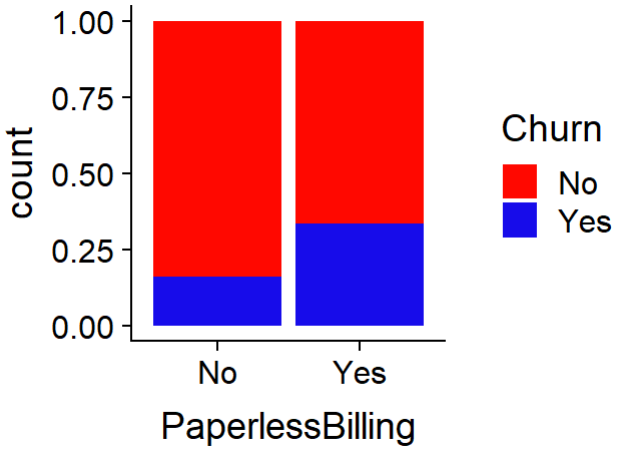
- InternetService, el tipo de servicio de internet si lo tiene, hay una tendencia clara al abandono entre los clientes que tienen fibra óptica.
- OnlineSecurity, si el cliente tiene seguridad online, hay una tendencia clara al abandono entre los clientes que no tienen seguridad online.
- OnlineBackup, si el cliente tiene backup online, se ve una tendencia al abandono aquellos que no tienen el servicio. +DeviceProtection, si el cliente tiene protección para el dispositivo, hay una ligera tendendica al abandono en aquellos que no tienen este servicio.
- TechSupport, si el cliente tiene soporte técnico, hay una tendencia clara al abandono entre los clientes que no tienen el servicio técnico.

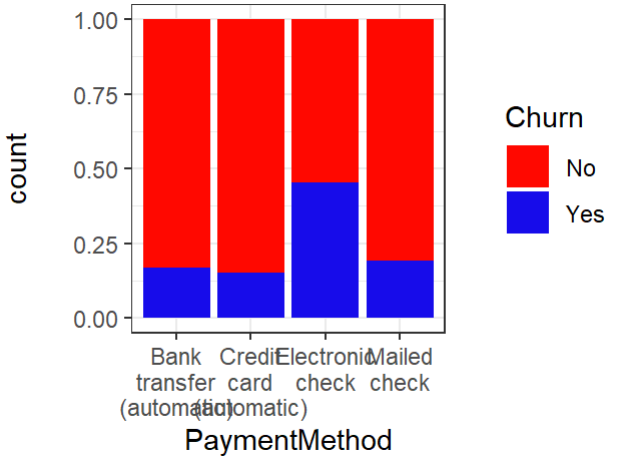
```
options(repr.plot.width = 12, repr.plot.height = 100)
plot_grid(ggplot(dataset, aes(x=StreamingTV,fill=Churn))+ geom_bar(position = 'fill') + scale
_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=Contract,fill=Churn))+ geom_bar(position = 'fill') + scale_fi
ll_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=PaperlessBilling,fill=Churn))+ geom_bar(position = 'fill') +
 scale_fill_manual(values=c("#FF0800", "#170CEA")),
          ggplot(dataset, aes(x=PaymentMethod,fill=Churn))+ geom_bar(position = 'fill') + sca
le_fill_manual(values=c("#FF0800", "#170CEA"))

            + theme_bw()+
          scale_x_discrete(labels = function(x) str_wrap(x, width = 10)),
          align = 'v', ncol=2)
```









En estas últimas variables hay una tendencia en los valores de cada variable según la variable objetivo Churn:

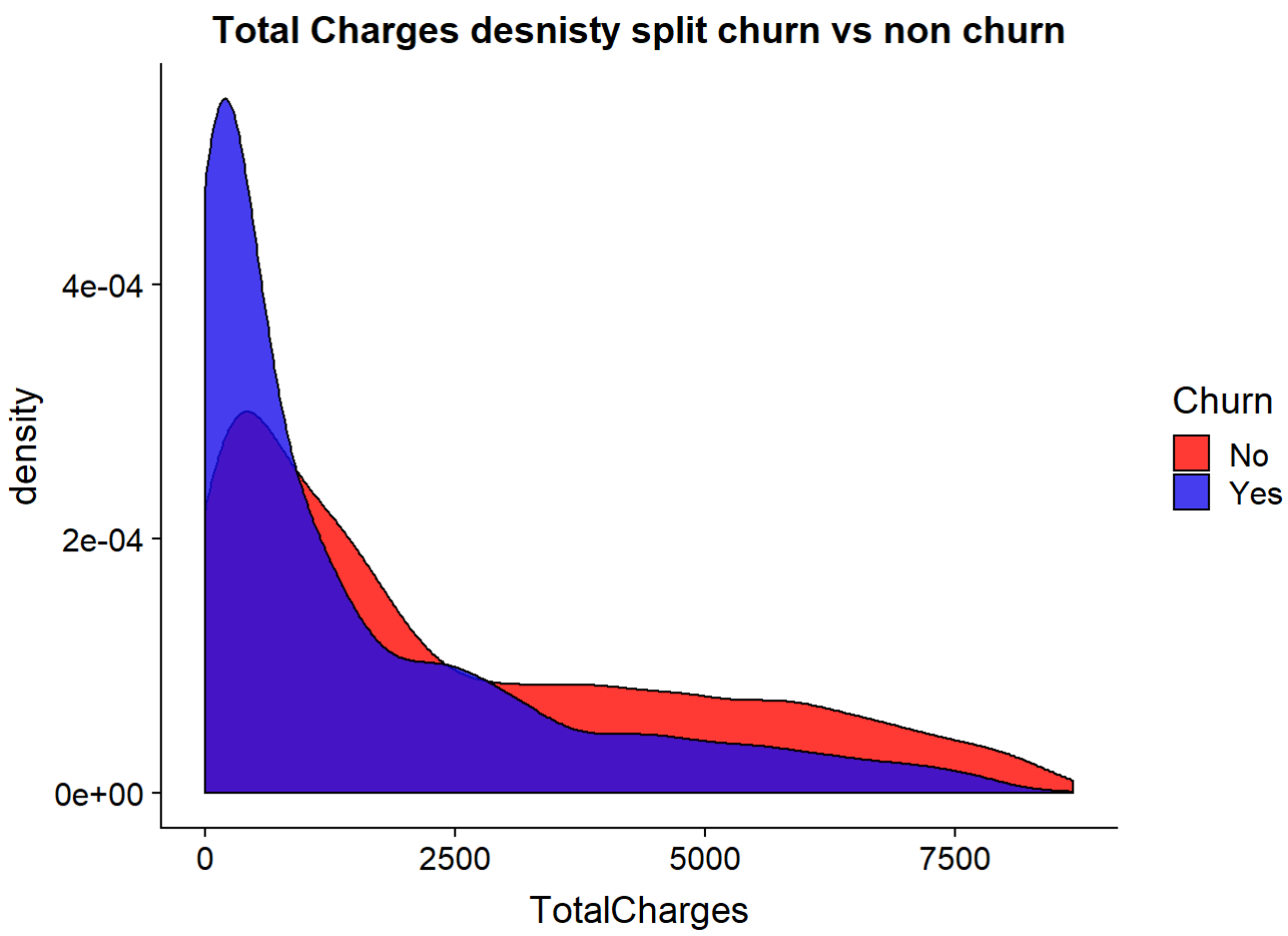
- StreamingTV, si el cliente servicio de TV Streamng, hay una ligera tendencia al abandono entre los clientes que tienen el servicio de TV Streaming y entiendiendo que se les puede sumar aquellos que no tienen servicio de internet ya que no disponen del servicio.
- Contract, el tipo de contrato que tienen los clientes, hay una tendencia muy clara entre los clientes que tienen un contrato de mes a mes.
- PaperlessBilling, si el cliente tiene factura digital, y hay una tendencia clara al abandono en los clientes que si tienen factura digital.
- PaymentMethod, el metodo de pago del cliente, hay una tendencia clara al abandono entre los clientes que tienen de metodo de pago con cheque electronico.

Variables continuas

Se va a tratar de analizar del mismo modo, el comportamiento de las variables continuas con respecto la variable de abandono, Churn:

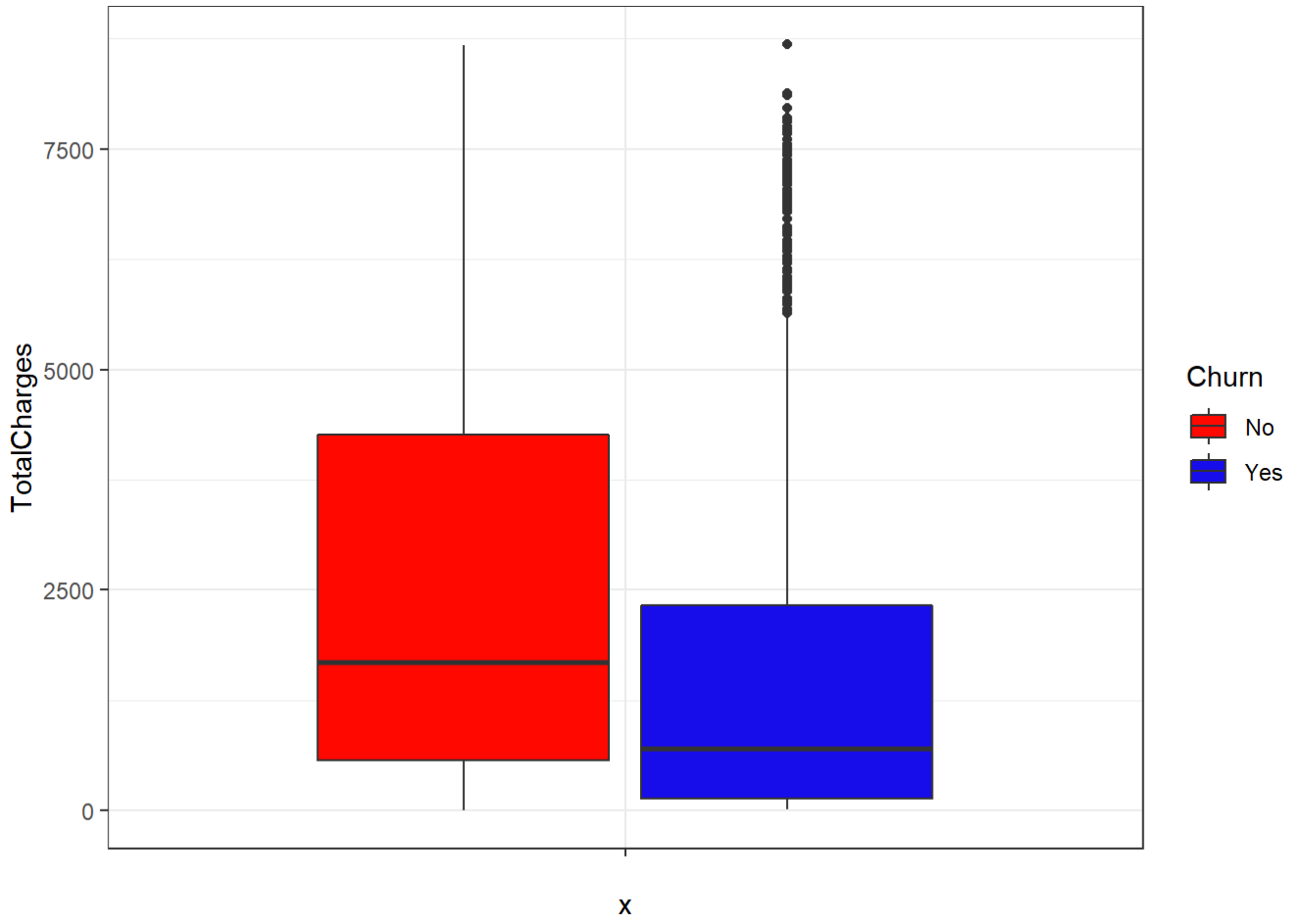
Disposicion de la variable TotalCharges contra la variable objetivo:

```
dataset %>% ggplot(aes(x=TotalCharges,fill=Churn))+ geom_density(alpha=0.8)+scale_fill_manual
(values=c("#FF0800", "#170CEA"))+labs(title='Total Charges desnisty split churn vs non churn'
)
```



Como se puede comprobar con este gráfico podemos observar que los usuarios cuanto menos tienen acumulado de pago más tienden al abandono.

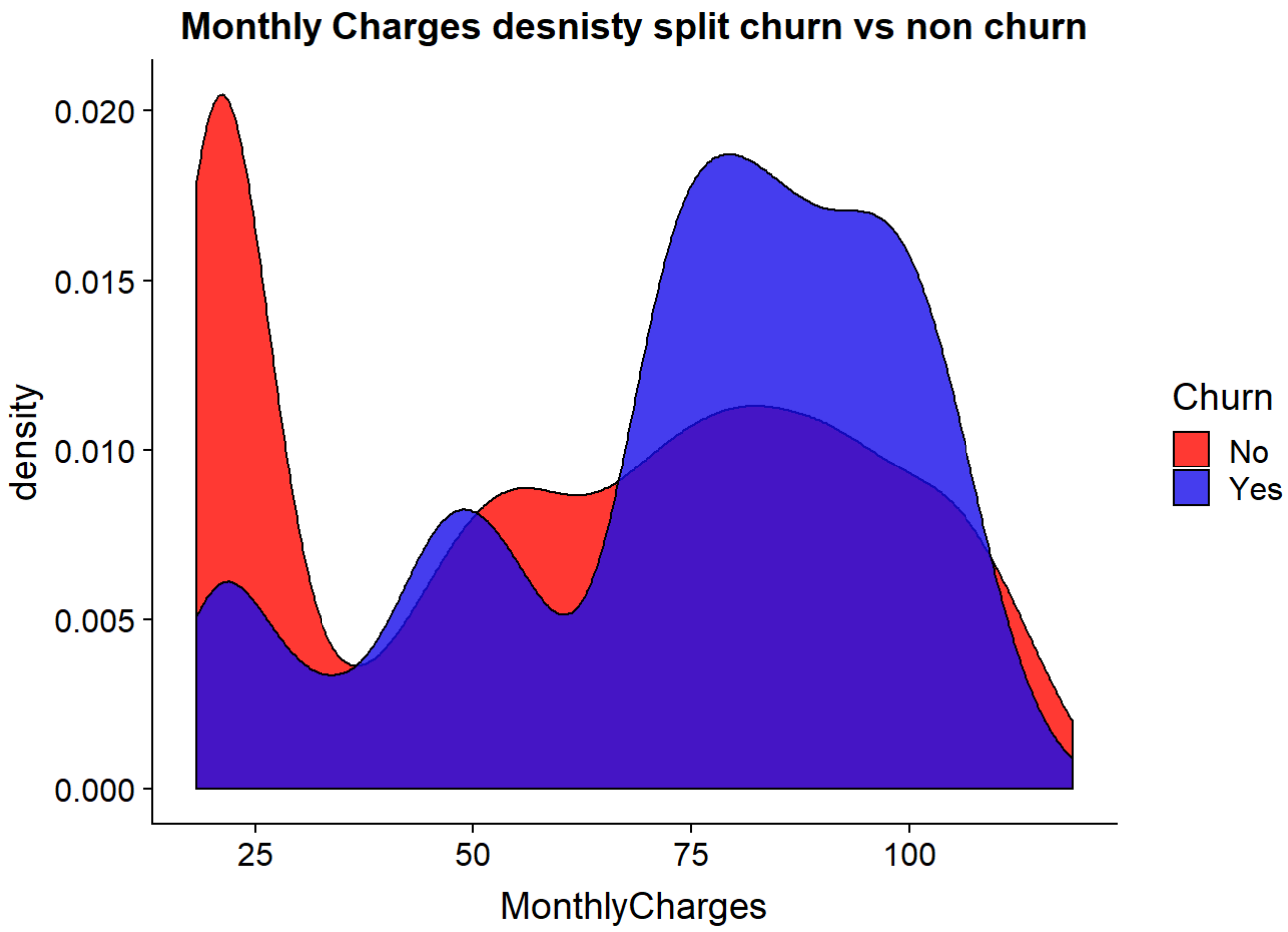
```
options(repr.plot.width = 6, repr.plot.height = 2)
ggplot(dataset, aes(y=TotalCharges, x=" ", fill=Churn))+
  scale_fill_manual(values=c("#FF0800", "#170CEA"))+
  geom_boxplot()+
  theme_bw()
```



Se puede confirmar lo visto en el gráfico anterior donde se ve una diferencia clara entre los segundos quartiles o medias de los que si y no abandonan. Viendo que hay una tendencia al abandono en aquellos que llevan menos cargos acumulados. Aunque puede verse algunos outliers, no tiene mucho aspecto de serlo, más bien pueden ser usuarios que lleven mucho tiempo ya en la compañia con cargos acumulados y que tiendan a hacer un cambio.

Disposicion de la variable MonthlyCharges contra la variable objetivo:

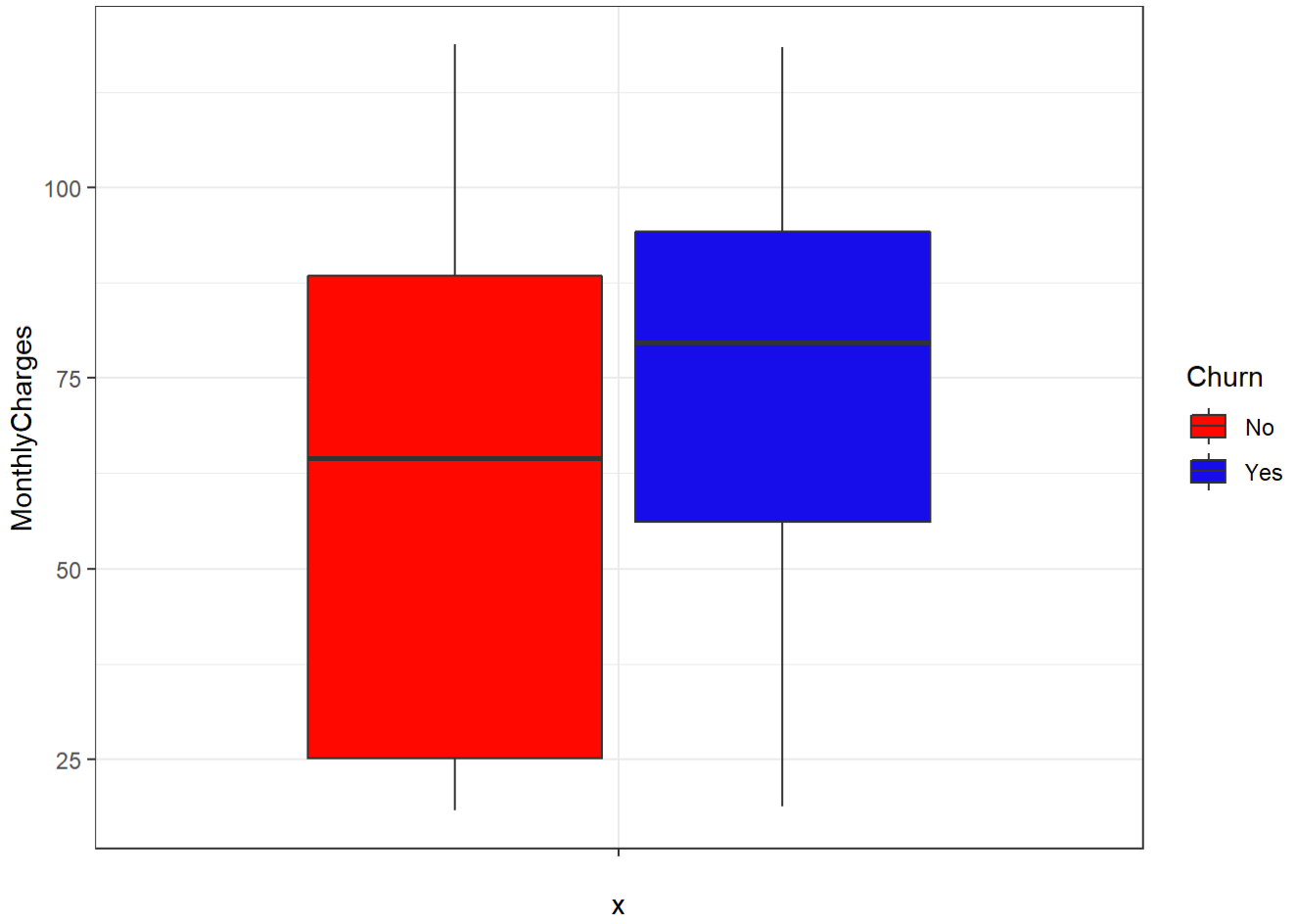
```
dataset %>% ggplot(aes(x=MonthlyCharges,fill=Churn))+ geom_density(alpha=0.8)+scale_fill_manu
al(values=c("#FF0800", "#170CEA"))+labs(title='Monthly Charges desnisty split churn vs non ch
urn' )
```



En cambio en el siguiente gráfico se puede observar que los que menos pagan mensualmente tienen a permanecer en la compañia mientras que los que más pagan tienen al abandono.

Boxplot de la variable MonthlyCharges sobre la objetivo Churn.

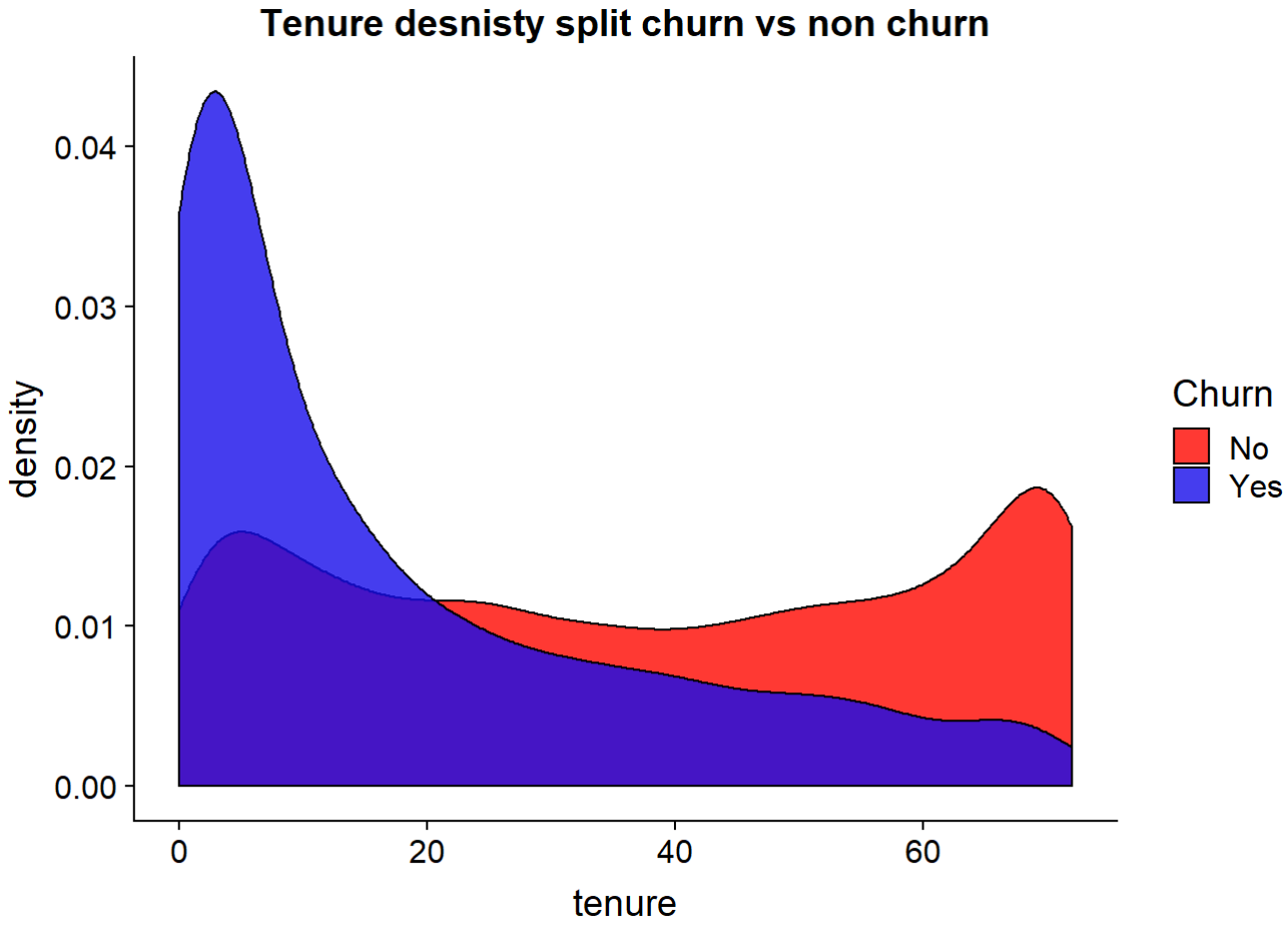
```
options(repr.plot.width = 6, repr.plot.height = 2)
ggplot(dataset, aes(y=MonthlyCharges, x=" ", fill=Churn))+
  scale_fill_manual(values=c("#FF0800", "#170CEA"))+
  geom_boxplot()+
  theme_bw()
```



Se ve una distrubución con valores altos más compactos en los que tienen tendencia al abandono contra los que no tienen, que es una distibución donde el primer quartil tiene un valor mucho más bajo y el segundo esta más parejo. Esto puede entenderse como que los que tienen tendencia al abandono tienen unos cargos mensuales altos frente a los que no lo tienen que esta algo más distribuido aunque con unos valores más bajos.

Distribución de la variable tenure:

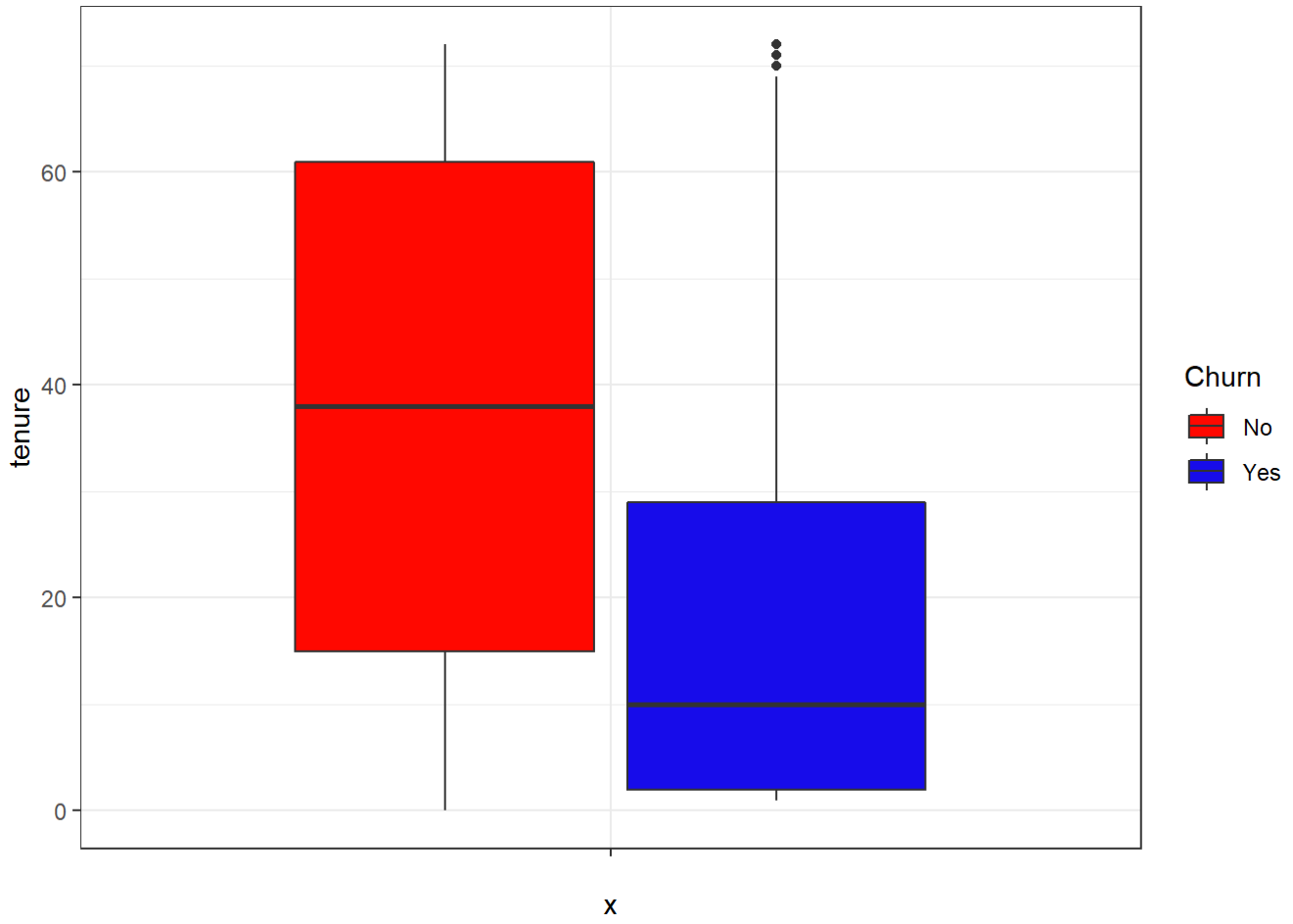
```
dataset %>% ggplot(aes(x=tenure,fill=Churn))+ geom_density(alpha=0.8)+scale_fill_manual(value
s=c("#FF0800", "#170CEA"))+labs(title='Tenure desnisty split churn vs non churn' )
```



Con el gráfico que puede observarse arriba se ve la relación entre el tiempo que lleva el usuario en la compañia y el abandono. Donde hay una tendencia clara a cuanto menos tiempo lleven en la compañia más abandono hay. Mientras que a medida que van avanzando en el tiempo de estancia en la compañia la tendencia del abandono tiende a menos.

Boxplot de la variable Tenure sobre la objetivo Churn.

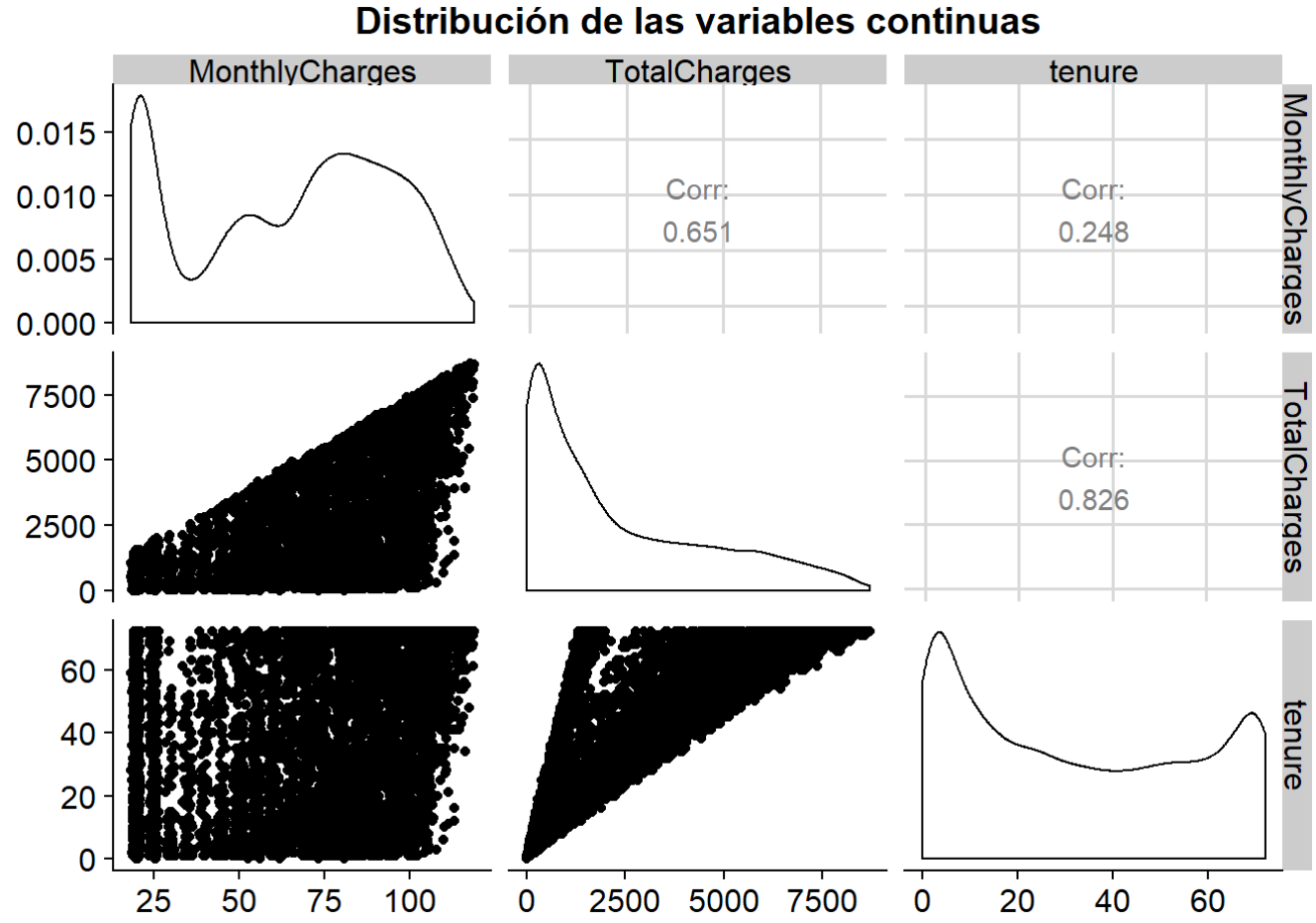
```
options(repr.plot.width = 6, repr.plot.height = 2)
ggplot(dataset, aes(y=tenure, x=" ", fill=Churn))+
  scale_fill_manual(values=c("#FF0800", "#170CEA"))+
  geom_boxplot()+
  theme_bw()
```



Se ve una diferencia bastante notable entre los usuarios que tienen más tendencia al abandono con los que no. Puede deducirse que los usuarios con mayor tendencia al abandono tienen unos valores de tenure bastante más bajos ya que el segundo quartil de los de no abandono esta en 40 meses frente a los que si con 10 meses. Esto confirma lo visto en la anterior gráfica donde se intuia que los usuarios con mayor tendencia al abandono son los que menos tiempo llevan en la compañia.

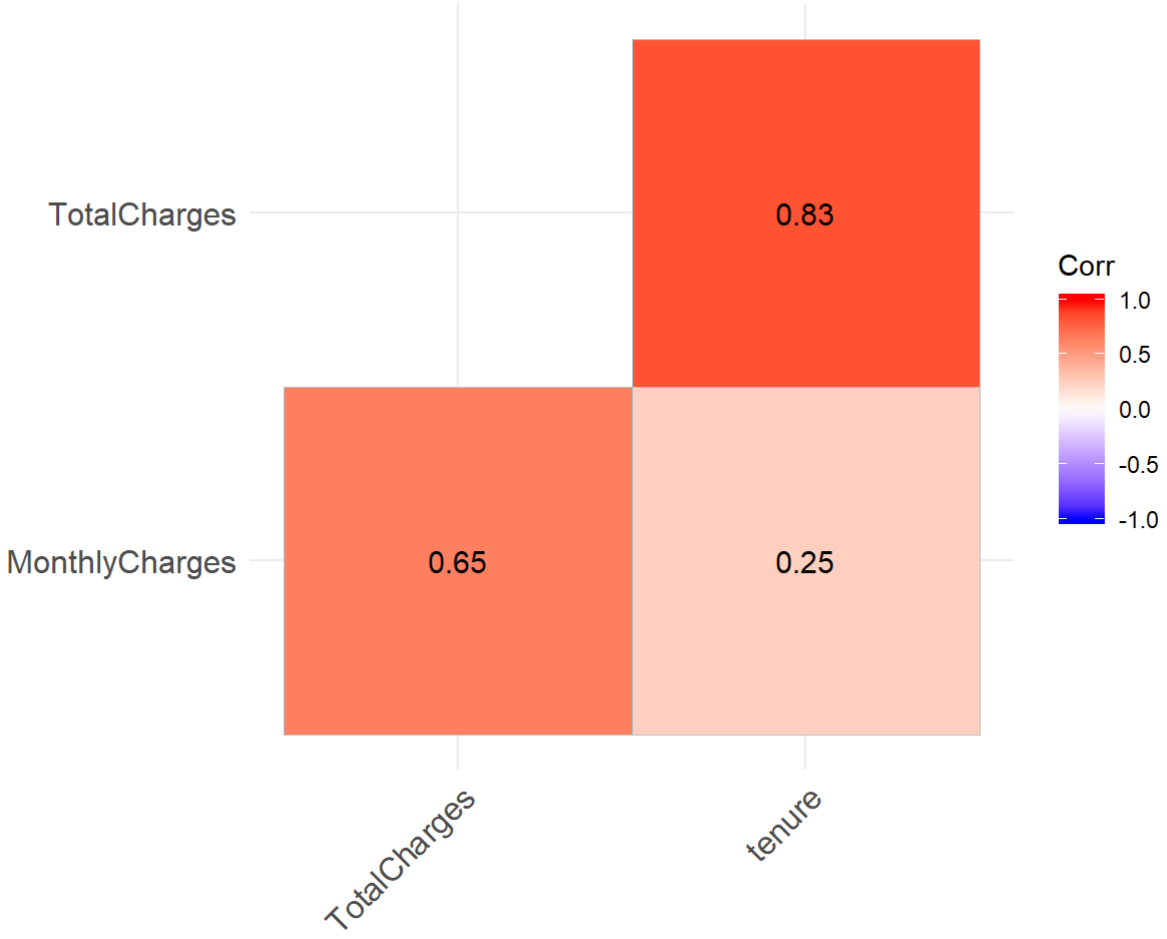
Distribución de las variables en scatterplot onde se va a observar alguna dependencia o correlación entre ellas y detección de outliers.

```
ggpairs(var_numeric, title = "Distribución de las variables continuas")
```



Tras observar posibles correlaciones entre las variables continuas, se ve necesario ver la matriz de correlaciones para poder definir las posibles dependencias que se empiezan a observar en el gráfico de distribución de las variables continuas.

```
corr <- cor(var_numeric,  method = "pearson", use = "complete.obs")
ggcorrplot(corr, hc.order = TRUE, type = "lower",
   lab = TRUE)
```



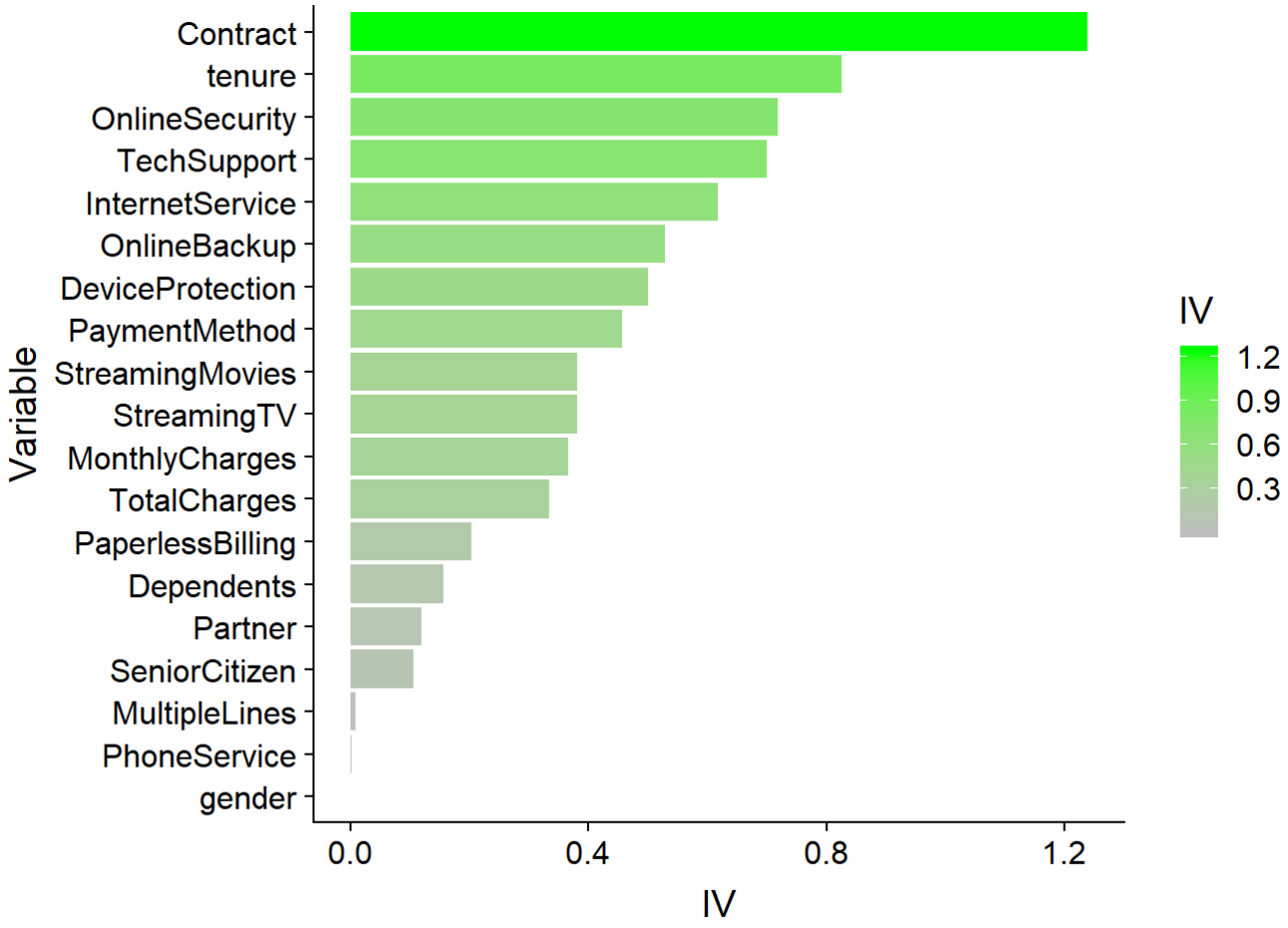
Tras analizar la matriz de correlaciones, puede observarse una alta dependencia positiva de la variable TotalCharges con Tenure y una algo más moderada con la otra variable MonthlyCharges.

Information value:

```
dataset_inf <- dataset
dataset_inf$Churn <- as.numeric(ifelse(dataset_inf$Churn=='Yes', 1, 0))
iv_ds <- create_infotables(data=dataset_inf,
                           y="Churn")
```

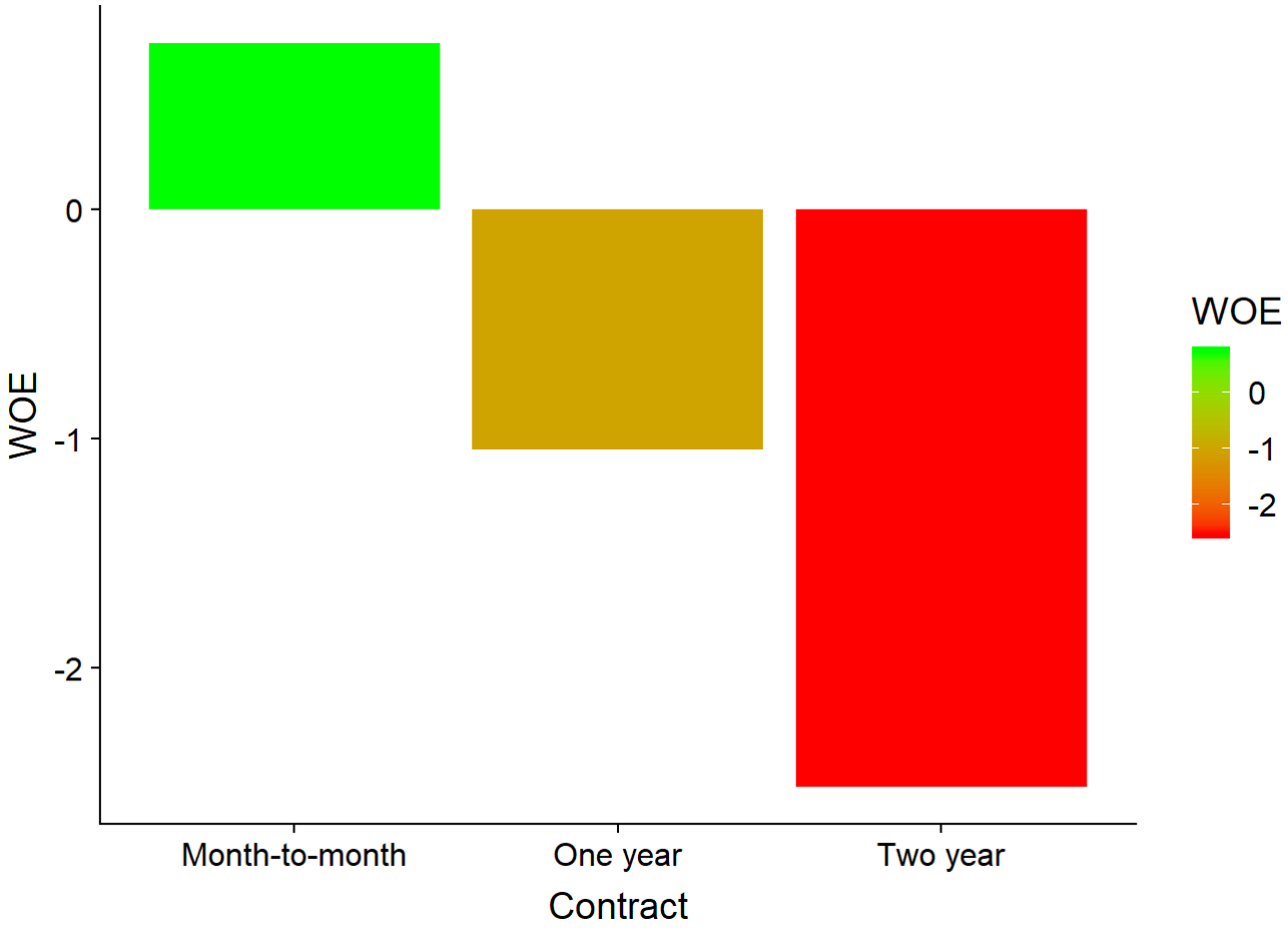
```
## [1] "Variable customerID was removed because it is a non-numeric variable with >1000 categ
ories"
```

```
iv_summary <- iv_ds$Summary
iv_summary <- iv_summary[order(iv_summary$IV), ]
iv_summary$Variable <- factor(iv_summary$Variable, levels=iv_summary$Variable)
ggplot(iv_summary, aes(x=Variable, y=IV, fill = IV))+
  coord_flip() +
  scale_fill_gradient(low = "grey", high = "green") +
  geom_bar(stat = "identity")
```



Ahora a verse el indicador de WOE para las tres principales variables vistas en los datos del information value. WOE segun la Contract:

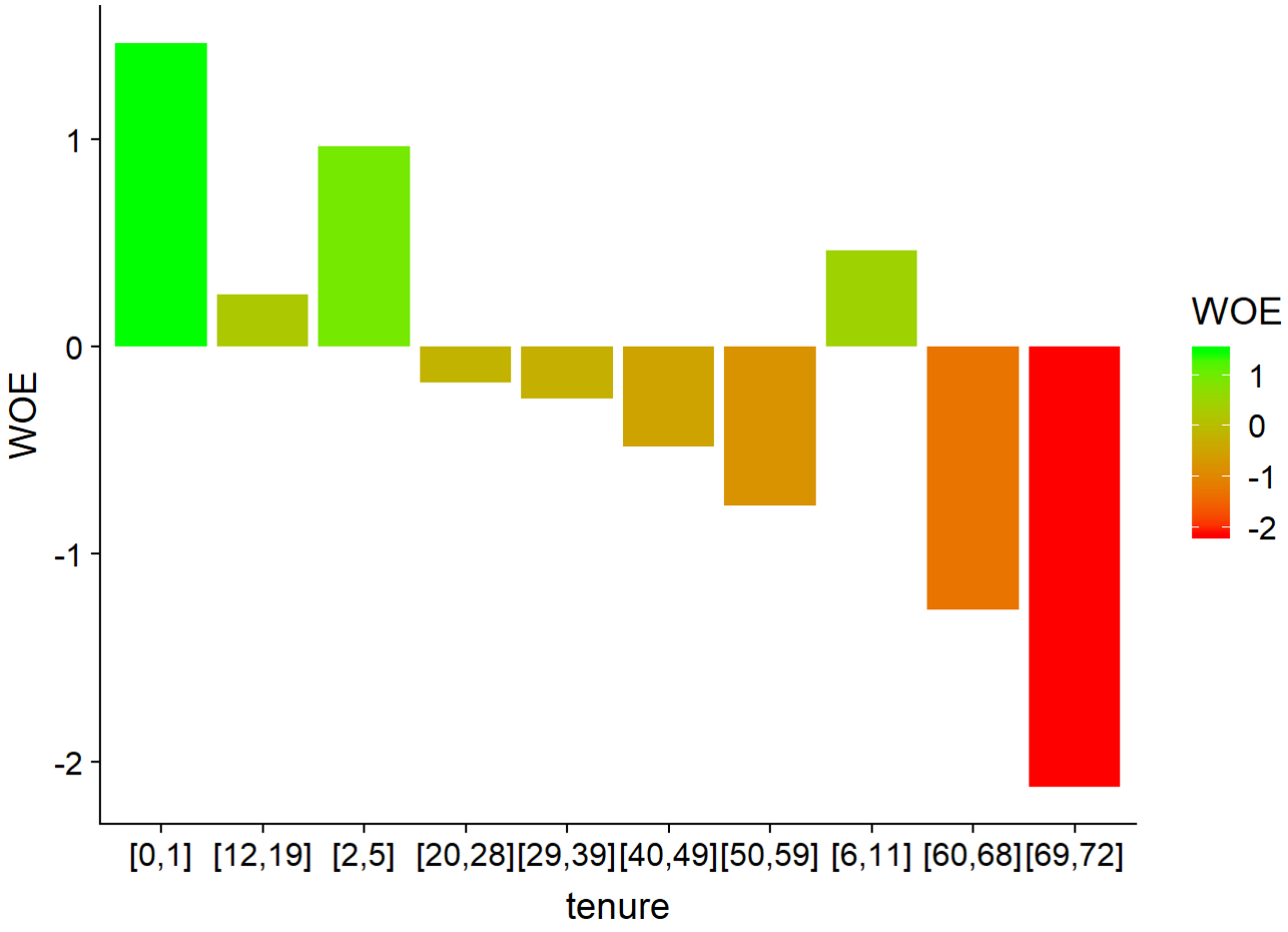
```
ggplot(iv_ds$Tables$Contract, aes(x=Contract, y=WOE, fill = WOE))+
  scale_fill_gradient(low = "red", high = "green") +
  geom_bar(stat = "identity")
```



Se empieza a confirmar lo visto anteriormente con una disposición al abandono aquellos que tienen un tipo de contrato de mes a mes frente al resto.

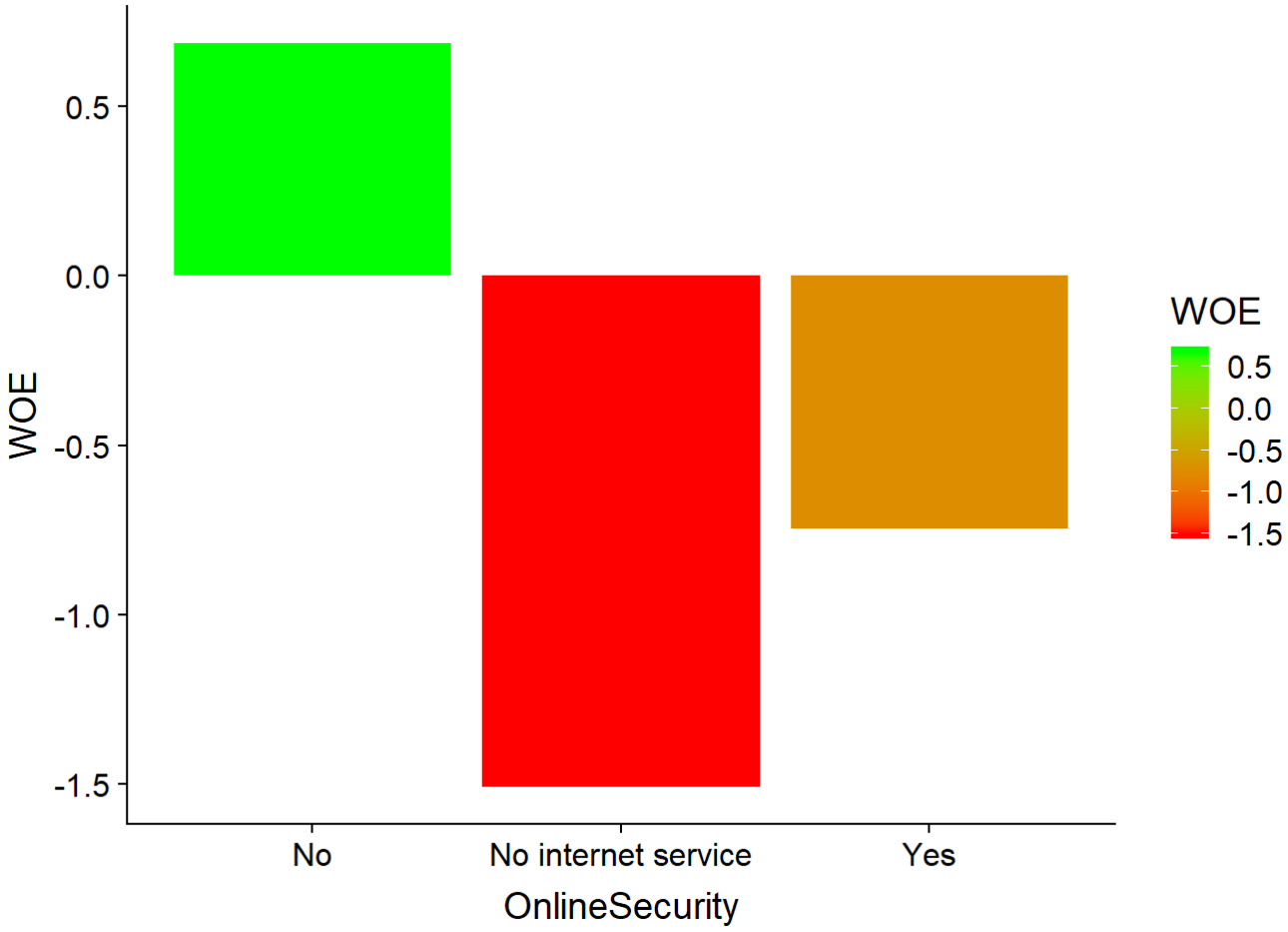
WOE según tenure:

```
ggplot(iv_ds$Tables$tenure, aes(x=tenure, y=WOE, fill = WOE))+
  scale_fill_gradient(low = "red", high = "green") +
  geom_bar(stat = "identity")
```



Aquí también se confirma lo visto anteriormente que hay una tendencia importante al abandono en aquellos que llevan menos menos tiempo en la compañia.

```
ggplot(iv_ds$Tables$OnlineSecurity, aes(x=OnlineSecurity, y=WOE, fill = WOE))+
  scale_fill_gradient(low = "red", high = "green") +
  geom_bar(stat = "identity")
```



En si los usuarios tienen servicio de seguridad online, se confirma una tendencia al abandono entre aquellos que no disponen de ese servicio.

Conclusiones EDA

Tras analizar todas y cada una de las variables y sus ditribuciones ya se ve un perfil claro el cual tiene la tendencia al abandono donde toman parte con más fuerza unas variables frente a otras.

Es un cliente que como bien se ve con las variables tenure, contract o totalcharges, lleva poco tiempo en la compañia lo cual tiene un cargo acumulado menor, tiene un contrato que se renueva mes a mes y lo que esto implica también un cargo mensual más alto.

Además se ve que sus condiciones son bastante básicas ya que tiende a carecer de servicios adicionales como el de OnlineSecurity, StreamingTV, DeviceProtection, OnlineBackup, TechSupport. Es un perfil de cliente que no suele tener Parners o Dependencies y que suele pagar con Electronic check.

Con todo esto puede resumirse en un perfil de cliente que busca un servicio de fibra óptica barato, sin ningún compromiso y tiende a ir de compañia en compañia haciendo pruebas de menos de año y medio sin muchas ataduras.

Modelado de los algoritmos

Instalación de librerias de depencias y prevencion/solventar errores

```
# Installation of the doSNOW parallel library with all dependencies
doInstall <- TRUE # Change to FALSE if you don't want packages installed.
toInstall <- c("doSNOW")
if((doInstall) && (!is.element(toInstall, installed.packages()[,1])))
{
    cat("Please install required package. Select server:"); chooseCRANmirror();
    install.packages(toInstall, dependencies = c("Depends", "Imports"))
}

# load doSnow and (parallel for CPU info) library
library(doSNOW)
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loading required package: iterators
```

```
## Loading required package: snow
```

```
library(parallel)
```

```
##
## Attaching package: 'parallel'
```

```
## The following objects are masked from 'package:snow':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, clusterSplit, makeCluster,
##     parApply, parCapply, parLapply, parRapply, parSapply,
##     splitIndices, stopCluster
```

```
# For doSNOW one can increase up to 128 nodes
# Each node requires 44 Mbyte RAM under WINDOWS.

# detect cores with parallel() package
nCores <- detectCores(logical = FALSE)
cat(nCores, " cores detected.")
```

```
## 4  cores detected.
```

```
# detect threads with parallel()
nThreads<- detectCores(logical = TRUE)
cat(nThreads, " threads detected.")
```

```
## 8  threads detected.
```

```
# Create doSNOW compute cluster (try 64)
# One can increase up to 128 nodes
# Each node requires 44 Mbyte RAM under WINDOWS.
cluster = makeCluster(nThreads, type = "SOCK")
class(cluster);
```

```
## [1] "SOCKcluster" "cluster"
```

```
# register the cluster
registerDoSNOW(cluster)

#get info
getDoParWorkers(); getDoParName();
```

```
## [1] 8
```

```
## [1] "doSNOW"
```

```
# insert parallel computation here

# stop cluster and remove clients
stopCluster(cluster); print("Cluster stopped.")
```

```
## [1] "Cluster stopped."
```

```
# insert serial backend, otherwise error in repetetive tasks
registerDoSEQ()

# clean up a bit.
invisible(gc); remove(nCores); remove(nThreads); remove(cluster);
```

Featuring engineering

Se ha visto que algunas variables categoricas dan la misma información en valores diferentes, como: ‘No’ y ‘No Internet Service’ o ‘No Phone Service’ y “NO”. Por lo tanto se van a reducir para no tener variables con la misma información en dos variables, cuando se pasen a dummies.

```
dataset_featured <- dataset
dataset_featured <- data.frame(lapply(dataset_featured, function(x) {
                  gsub("No internet service", "No", x)}))

dataset_featured <- data.frame(lapply(dataset_featured, function(x) {
                  gsub("No phone service", "No", x)}))
```

En cuanto a las variables continuas, primero se va a confirmar y estandarizar que todas son numéricas.

```
num_columns <- c("tenure", "MonthlyCharges", "TotalCharges")
dataset_featured[num_columns] <- sapply(dataset_featured[num_columns], as.numeric)
```

A continuación se a escalar las variables continuas para poder tratar mejor con ellas en los modelos y que esten bajo el mismo rango a la hora de que el modelo pueda entender el peso de cada una.

```
dataset_featured$tenure <- scale(dataset_featured$tenure, scale = T)
dataset_featured$MonthlyCharges <- scale(dataset_featured$MonthlyCharges, scale = T)
dataset_featured$TotalCharges <- scale(dataset_featured$TotalCharges, scale = T)
```

Pasar las variables categoricas a dummies para poder trabajar con ellas en los modelos.

```
dataset_churn_factor <- dataset_featured
dataset_churn_factor$Churn <- as.factor(ifelse(dataset_churn_factor$Churn=='Yes', 1, 0))
data_model <- dataset_churn_factor[,!names(dataset_churn_factor) %in% c("customerID")]
var_categoric <- names(data_model[,!names(data_model) %in% c(var_num_total, "Churn")])
data_model_dummy <- dummy_cols(data_model, select_columns = var_categoric, remove_first_dummy
= TRUE)
data_model_dummy_only <- data_model_dummy[,!names(data_model_dummy) %in% var_categoric]
```

Dividir entre entrenamiento y test

```
train <- createDataPartition(data_model_dummy_only$Churn, p = 0.8, list = F)
data_train <- data_model_dummy_only[train,]
data_test <- data_model_dummy_only[-train,]
dim(data_train)
```

```
## [1] 5636   24
```

```
dim(data_test)
```

```
## [1] 1407   24
```

GLM

Antes de empezar mencionar como se entiende que se debiera valorar o que métricas debieran primar a la hora de evaluar los modelos. Como lo que se trata de detectar el mayor numero de clientes que abandonen la compañia y que no se escape ninguno, hay que tener más en cuenta la sensibilidad (ya que como valor positivo se va a tratar el 1 de abandono, ya que se quiere optimizar para su detección) junto con el auc, aunque se buscará un equilibrio de todo.

Por otro lado, a la hora de entrenar cada modelo, se va a hacer con el metodo de cross validation, que consiste en dividir la muestra por n y cada división hara una vez de test. Para entrenar con diferentes muestras y luego sacar un promedio de todos los resultados. De esta manera se eficienta el entrenaimiento, ya que se entrena varias veces con diferentes modelos. Además se ha decido tras algunas pruebas de concepto, que el dataset si que puede tener algo de desbalanceo, por lo que se ha seleccionado la tecnica de Oversampling, SMOTE. Este metodo selecciona dos instancias similares utilizando vecinos más cercanos y bootstrapping, y genera muestras sintéticas a partir de instancias de las clases minoritarias.

Se va a entrenar el modelo de GLM

```
set.seed(46)
# Definir train control para cross validation
train_control <- trainControl(method="cv", number=10, sampling="smote")
# Entrenar el modelo
model_glm_train <- train(Churn~., data=data_train, trControl=train_control, method="glm")
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# Imprimir resultados
print(model_glm_train)
```

```
## Generalized Linear Model
##
## 5636 samples
##   23 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5073, 5072, 5072, 5072, 5073, 5072, ...
## Addtional sampling using SMOTE
##
## Resampling results:
##
##   Accuracy   Kappa
##   0.7634786  0.4560921
```

Una vez entrenado y viendo que el accuracy en entrenamiento es menor del resultado de H2O de AutoML, se va a comprobar su eficiencia predictora y a hacer la matriz de confusión.

```
model_glm_predict <- predict(model_glm_train, data_test)
confusionMatrix(model_glm_predict, data_test$Churn, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 807 108
##          1 227 265
##
##                Accuracy : 0.7619
##                  95% CI : (0.7388, 0.7839)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 0.01114
##
##                   Kappa : 0.4455
##
##  Mcnemar's Test P-Value : 1.141e-10
##
##             Sensitivity : 0.7105
##             Specificity : 0.7805
##          Pos Pred Value : 0.5386
##          Neg Pred Value : 0.8820
##              Prevalence : 0.2651
##          Detection Rate : 0.1883
##    Detection Prevalence : 0.3497
##       Balanced Accuracy : 0.7455
##
##        'Positive' Class : 1
##
```

Por lo tanto este algoritmo tiene un buen auc con una sensiblidad y especificidad buenas y equilibradas. Detecta bastante bien los verdaderos positivos en abandono y los falsos negativos.

GBM

Se va a entrenar el modelo seleccionado por H20.AutoML como mejor.

```
set.seed(46)
# Definir train control para cross validation
train_control <- trainControl(method="cv", number=10, sampling = "smote")
# Entrenar el modelo
model_gbm_train <- train(Churn~., data=data_train, trControl=train_control, method="gbm", ver
bose = FALSE)
# Imprimir resultados
print(model_gbm_train)
```

```
## Stochastic Gradient Boosting
##
## 5636 samples
##   23 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5073, 5072, 5072, 5072, 5073, 5072, ...
## Addtional sampling using SMOTE
##
## Resampling results across tuning parameters:
##
##   interaction.depth  n.trees  Accuracy   Kappa
##   1                   50      0.7657833  0.4522352
##   1                  100      0.7728827  0.4674467
##   1                  150      0.7751890  0.4704180
##   2                   50      0.7771399  0.4642230
##   2                  100      0.7810454  0.4601084
##   2                  150      0.7884975  0.4669338
##   3                   50      0.7819297  0.4679435
##   3                  100      0.7858376  0.4617550
##   3                  150      0.7881436  0.4585598
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were n.trees = 150,
##  interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10.
```

En este se indica la mejor configuración del set de entrenamiento para el modelo de GBM. Esto lleva a un auc que sigue siendo menor al que ha salido con H2o. Ahora va a procederse a predecir con el test y ver la matriz de confusión que tal se comporta.

```
model_gbm_predict <- predict(model_gbm_train,data_test)
confusionMatrix(model_gbm_predict,data_test$Churn, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 881 143
##          1 153 230
##
##                Accuracy : 0.7896
##                  95% CI : (0.7674, 0.8107)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 1.127e-06
##
##                   Kappa : 0.4647
##
##  Mcnemar's Test P-Value : 0.6009
##
##             Sensitivity : 0.6166
##             Specificity : 0.8520
##          Pos Pred Value : 0.6005
##          Neg Pred Value : 0.8604
##              Prevalence : 0.2651
##          Detection Rate : 0.1635
##    Detection Prevalence : 0.2722
##       Balanced Accuracy : 0.7343
##
##        'Positive' Class : 1
##
```

Este modelo da un auc algo mas alto, aunque la sensibilidad es algo más baja ya que hay más falsos negativos que en el anterior modelo. Como se ha comentado, aunque la idea es optimizar todos los indicadores, prima el poder detectar todos los que realmente son positivos.

XGBoost

```
set.seed(46)
# Definir train control para cross validation
train_control <- trainControl(method="cv", number=10, sampling = "smote")
# Entrenar el modelo
model_xgb_train <- train(Churn~., data=data_train, trControl=train_control, method="xgbTree",
verbose = FALSE)
# Imprimir resultados
print(model_xgb_train)
```

```
## eXtreme Gradient Boosting
##
## 5636 samples
##   23 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5073, 5072, 5072, 5072, 5073, 5072, ...
## Addtional sampling using SMOTE
##
## Resampling results across tuning parameters:
##
##   eta  max_depth  colsample_bytree  subsample  nrounds  Accuracy
##   0.3  1          0.6               0.50        50      0.7704020
##   0.3  1          0.6               0.50       100      0.7721741
##   0.3  1          0.6               0.50       150      0.7709311
##   0.3  1          0.6               0.75        50      0.7703989
##   0.3  1          0.6               0.75       100      0.7682690
##   0.3  1          0.6               0.75       150      0.7716425
##   0.3  1          0.6               1.00        50      0.7689795
##   0.3  1          0.6               1.00       100      0.7695127
##   0.3  1          0.6               1.00       150      0.7693347
##   0.3  1          0.8               0.50        50      0.7725281
##   0.3  1          0.8               0.50       100      0.7719946
##   0.3  1          0.8               0.50       150      0.7727039
##   0.3  1          0.8               0.75        50      0.7711062
##   0.3  1          0.8               0.75       100      0.7721723
##   0.3  1          0.8               0.75       150      0.7714627
##   0.3  1          0.8               1.00        50      0.7728837
##   0.3  1          0.8               1.00       100      0.7732389
##   0.3  1          0.8               1.00       150      0.7760777
##   0.3  2          0.6               0.50        50      0.7861929
##   0.3  2          0.6               0.50       100      0.7899201
##   0.3  2          0.6               0.50       150      0.7861919
##   0.3  2          0.6               0.75        50      0.7824692
##   0.3  2          0.6               0.75       100      0.7808706
##   0.3  2          0.6               0.75       150      0.7830017
##   0.3  2          0.6               1.00        50      0.7803387
##   0.3  2          0.6               1.00       100      0.7819363
##   0.3  2          0.6               1.00       150      0.7824692
##   0.3  2          0.8               0.50        50      0.7829973
##   0.3  2          0.8               0.50       100      0.7833510
##   0.3  2          0.8               0.50       150      0.7792723
##   0.3  2          0.8               0.75        50      0.7831749
##   0.3  2          0.8               0.75       100      0.7801607
##   0.3  2          0.8               0.75       150      0.7817552
##   0.3  2          0.8               1.00        50      0.7835273
##   0.3  2          0.8               1.00       100      0.7833522
##   0.3  2          0.8               1.00       150      0.7842391
##   0.3  3          0.6               0.50        50      0.7829998
##   0.3  3          0.6               0.50       100      0.7748384
##   0.3  3          0.6               0.50       150      0.7746605
##   0.3  3          0.6               0.75        50      0.7847722
##   0.3  3          0.6               0.75       100      0.7803368
##   0.3  3          0.6               0.75       150      0.7766115
##   0.3  3          0.6               1.00        50      0.7849540
```

```
##   0.3  3          0.6              1.00       100      0.7805147
##   0.3  3          0.6              1.00       150      0.7812227
##   0.3  3          0.8              0.50        50      0.7812239
##   0.3  3          0.8              0.50       100      0.7803393
##   0.3  3          0.8              0.50       150      0.7721767
##   0.3  3          0.8              0.75        50      0.7821139
##   0.3  3          0.8              0.75       100      0.7787388
##   0.3  3          0.8              0.75       150      0.7766105
##   0.3  3          0.8              1.00        50      0.7806923
##   0.3  3          0.8              1.00       100      0.7780290
##   0.3  3          0.8              1.00       150      0.7783845
##   0.4  1          0.6              0.50        50      0.7705752
##   0.4  1          0.6              0.50       100      0.7716425
##   0.4  1          0.6              0.50       150      0.7735945
##   0.4  1          0.6              0.75        50      0.7686239
##   0.4  1          0.6              0.75       100      0.7705756
##   0.4  1          0.6              0.75       150      0.7750110
##   0.4  1          0.6              1.00        50      0.7727035
##   0.4  1          0.6              1.00       100      0.7712829
##   0.4  1          0.6              1.00       150      0.7719934
##   0.4  1          0.8              0.50        50      0.7705759
##   0.4  1          0.8              0.50       100      0.7709321
##   0.4  1          0.8              0.50       150      0.7759007
##   0.4  1          0.8              0.75        50      0.7744794
##   0.4  1          0.8              0.75       100      0.7730607
##   0.4  1          0.8              0.75       150      0.7766121
##   0.4  1          0.8              1.00        50      0.7730591
##   0.4  1          0.8              1.00       100      0.7735932
##   0.4  1          0.8              1.00       150      0.7751890
##   0.4  2          0.6              0.50        50      0.7792726
##   0.4  2          0.6              0.50       100      0.7801592
##   0.4  2          0.6              0.50       150      0.7810466
##   0.4  2          0.6              0.75        50      0.7867235
##   0.4  2          0.6              0.75       100      0.7798071
##   0.4  2          0.6              0.75       150      0.7824670
##   0.4  2          0.6              1.00        50      0.7819319
##   0.4  2          0.6              1.00       100      0.7826408
##   0.4  2          0.6              1.00       150      0.7861904
##   0.4  2          0.8              0.50        50      0.7812202
##   0.4  2          0.8              0.50       100      0.7838810
##   0.4  2          0.8              0.50       150      0.7810429
##   0.4  2          0.8              0.75        50      0.7821064
##   0.4  2          0.8              0.75       100      0.7814009
##   0.4  2          0.8              0.75       150      0.7815782
##   0.4  2          0.8              1.00        50      0.7869024
##   0.4  2          0.8              1.00       100      0.7845949
##   0.4  2          0.8              1.00       150      0.7815801
##   0.4  3          0.6              0.50        50      0.7808718
##   0.4  3          0.6              0.50       100      0.7805138
##   0.4  3          0.6              0.50       150      0.7826480
##   0.4  3          0.6              0.75        50      0.7796266
##   0.4  3          0.6              0.75       100      0.7762537
##   0.4  3          0.6              0.75       150      0.7712838
##   0.4  3          0.6              1.00        50      0.7801573
##   0.4  3          0.6              1.00       100      0.7826468
##   0.4  3          0.6              1.00       150      0.7794531
##   0.4  3          0.8              0.50        50      0.7796263
##   0.4  3          0.8              0.50       100      0.7728853
```

```
##   0.4  3          0.8              0.50       150      0.7714674
##   0.4  3          0.8              0.75        50      0.7796250
##   0.4  3          0.8              0.75       100      0.7780325
##   0.4  3          0.8              0.75       150      0.7707566
##   0.4  3          0.8              1.00        50      0.7771437
##   0.4  3          0.8              1.00       100      0.7783871
##   0.4  3          0.8              1.00       150      0.7780375
##   Kappa
##   0.4644527
##   0.4598679
##   0.4533849
##   0.4640945
##   0.4546085
##   0.4567877
##   0.4613002
##   0.4569791
##   0.4520587
##   0.4658611
##   0.4599420
##   0.4557719
##   0.4654685
##   0.4618677
##   0.4562825
##   0.4700101
##   0.4654624
##   0.4679726
##   0.4594463
##   0.4533938
##   0.4389926
##   0.4512478
##   0.4309008
##   0.4262814
##   0.4431891
##   0.4284761
##   0.4269322
##   0.4505115
##   0.4353795
##   0.4195877
##   0.4542634
##   0.4269800
##   0.4231168
##   0.4536333
##   0.4352239
##   0.4320773
##   0.4418051
##   0.4097322
##   0.4027783
##   0.4484745
##   0.4244613
##   0.4080281
##   0.4487831
##   0.4242484
##   0.4224399
##   0.4393593
##   0.4277137
##   0.4034913
##   0.4389645
##   0.4162329
```

```
##   0.4100365
##   0.4370362
##   0.4175747
##   0.4148632
##   0.4610501
##   0.4545650
##   0.4543146
##   0.4586014
##   0.4559174
##   0.4588619
##   0.4661125
##   0.4572987
##   0.4541309
##   0.4609586
##   0.4542000
##   0.4588905
##   0.4679500
##   0.4588276
##   0.4614270
##   0.4678817
##   0.4625950
##   0.4613800
##   0.4344318
##   0.4272491
##   0.4214809
##   0.4539822
##   0.4217689
##   0.4216440
##   0.4416911
##   0.4299060
##   0.4321757
##   0.4392816
##   0.4349512
##   0.4191443
##   0.4419409
##   0.4244613
##   0.4179799
##   0.4556579
##   0.4344352
##   0.4222201
##   0.4316994
##   0.4198790
##   0.4271560
##   0.4262284
##   0.4106374
##   0.3972378
##   0.4330360
##   0.4264459
##   0.4187792
##   0.4228731
##   0.4003664
##   0.3992265
##   0.4281739
##   0.4128092
##   0.3944331
##   0.4237014
##   0.4136073
##   0.4130000
```

```
##
## Tuning parameter 'gamma' was held constant at a value of 0
##
## Tuning parameter 'min_child_weight' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were nrounds = 100, max_depth = 2,
##  eta = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1
##  and subsample = 0.5.
```

Tras entrenar el modelo con XGBoost se encuentra una combinación de parametros donde como resultante el auc aún menor que el resultado de H20. Se va a proceder a predecir y ver la matriz de confusión para ver como se comporta el modelo.

```
model_xgb_predict <- predict(model_xgb_train,data_test)
confusionMatrix(model_xgb_predict,data_test$Churn, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 889 165
##          1 145 208
##
##                Accuracy : 0.7797
##                  95% CI : (0.7571, 0.8011)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 6.05e-05
##
##                   Kappa : 0.4247
##
##  Mcnemar's Test P-Value : 0.2805
##
##             Sensitivity : 0.5576
##             Specificity : 0.8598
##          Pos Pred Value : 0.5892
##          Neg Pred Value : 0.8435
##              Prevalence : 0.2651
##          Detection Rate : 0.1478
##    Detection Prevalence : 0.2509
##       Balanced Accuracy : 0.7087
##
##        'Positive' Class : 1
##
```

Este algoritmo finalmente tiene un auc similar al anterior aunque tiene una sensibilidad peor, ya que como puede observase en la matriz de confusión los falsos negativos son más altos aún que en los anteriores casos.

Random Forest

```
set.seed(46)
# Definir train control para cross validation
train_control <- trainControl(method="cv", number=10, sampling = "smote")
# Entrenar el modelo
model_rf_train <- train(Churn~., data=data_train, trControl=train_control, method="rf", verbo
se = FALSE)
# Imprimir resultados
print(model_rf_train)
```

```
## Random Forest
##
## 5636 samples
##   23 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5073, 5072, 5072, 5072, 5073, 5072, ...
## Addtional sampling using SMOTE
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.7838835  0.4519408
##   12    0.7629470  0.4127247
##   23    0.7510585  0.3894993
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

Tras el entrenamiento el algoritmo de Random Forest da un auc aún inferior al resultado obtenido en H2O. Se a a ver que tal se comporta prediciendo con la parte del test y viendo la matriz de confusión.

```
model_rf_predict <- predict(model_rf_train,data_test)
confusionMatrix(model_rf_predict,data_test$Churn, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 867 140
##          1 167 233
##
##                Accuracy : 0.7818
##                  95% CI : (0.7593, 0.8031)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 2.754e-05
##
##                   Kappa : 0.4527
##
##  Mcnemar's Test P-Value : 0.1378
##
##             Sensitivity : 0.6247
##             Specificity : 0.8385
##          Pos Pred Value : 0.5825
##          Neg Pred Value : 0.8610
##              Prevalence : 0.2651
##          Detection Rate : 0.1656
##    Detection Prevalence : 0.2843
##       Balanced Accuracy : 0.7316
##
##        'Positive' Class : 1
##
```

Este modelo tiene un auc de los más altos hasta ahora. No obstante aunque es el segundo algoritmo en verdaderos positivos (detectar la gente que va a abandonar), tiene una sensibilidad algo baja con respecto a algún otro modelo, lo que hace que este más alto los falsos negativos (gente que realmente abandona la compañia pero que el algoritmo predice que no van a abandonar).

GLMNET

```
set.seed(46)
# Definir train control para cross validation
train_control <- trainControl(method="cv", number=10, sampling = "smote")
# Entrenar el modelo
model_glmnet_train <- train(Churn~., data=data_train, trControl=train_control, method="glmne
t")
# Imprimir resultados
print(model_glmnet_train)
```

```
## glmnet
##
## 5636 samples
##   23 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5073, 5072, 5072, 5072, 5073, 5072, ...
## Addtional sampling using SMOTE
##
## Resampling results across tuning parameters:
##
##   alpha  lambda        Accuracy   Kappa
##   0.10   0.0002690401  0.7654290  0.4590286
##   0.10   0.0026904005  0.7654290  0.4588230
##   0.10   0.0269040051  0.7684466  0.4626031
##   0.55   0.0002690401  0.7627666  0.4530330
##   0.55   0.0026904005  0.7632982  0.4541399
##   0.55   0.0269040051  0.7617024  0.4513930
##   1.00   0.0002690401  0.7625890  0.4528235
##   1.00   0.0026904005  0.7629458  0.4547919
##   1.00   0.0269040051  0.7574440  0.4449997
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0.1 and lambda
##  = 0.02690401.
```

Tras ver que el modelo de GLMNET con el set de entrenamiento da que tampoco supera el modelo de H20 con las parametrizaciones optimizadas, se va a comprobar como se comporta prediciendo y en la matriz de confusión.

```
model_glmnet_predict <- predict(model_glmnet_train,data_test)
confusionMatrix(model_glmnet_predict,data_test$Churn, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 817 112
##          1 217 261
##
##                Accuracy : 0.7662
##                  95% CI : (0.7432, 0.7881)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 0.003939
##
##                   Kappa : 0.4494
##
##  Mcnemar's Test P-Value : 9.826e-09
##
##             Sensitivity : 0.6997
##             Specificity : 0.7901
##          Pos Pred Value : 0.5460
##          Neg Pred Value : 0.8794
##              Prevalence : 0.2651
##          Detection Rate : 0.1855
##    Detection Prevalence : 0.3397
##       Balanced Accuracy : 0.7449
##
##        'Positive' Class : 1
##
```

Este modelo tiene un auc algo por debajo de algunos pero con una sensibilidad mayor y una especificidad no tan alta como otros. Aún así no es el algoritmo con la sensibilidad, especificidad y auc más equilibradas pero no por mucho. Es el segundo algoritmo con mayor acierto en verdaderos positivos y que menos tiene en falsos negativos.

Conclusiones del modelado

Después de varias pruebas de concepto:

- probando con diferentes featuring egineering (categorizar tenure, escalar de modo continua, agregar valores de “NO” y “No internet service”,…) y sin ellas,
- probando sin y con oversampling (down y SMOTE),
- usando el paquete caret que prueba con diferentes hiperparamtros en el entrenamiento para cada algoritmo y les asigna la mejor combinación de ellos,
- probar con varios algoritmos, confunciones de cross validation,

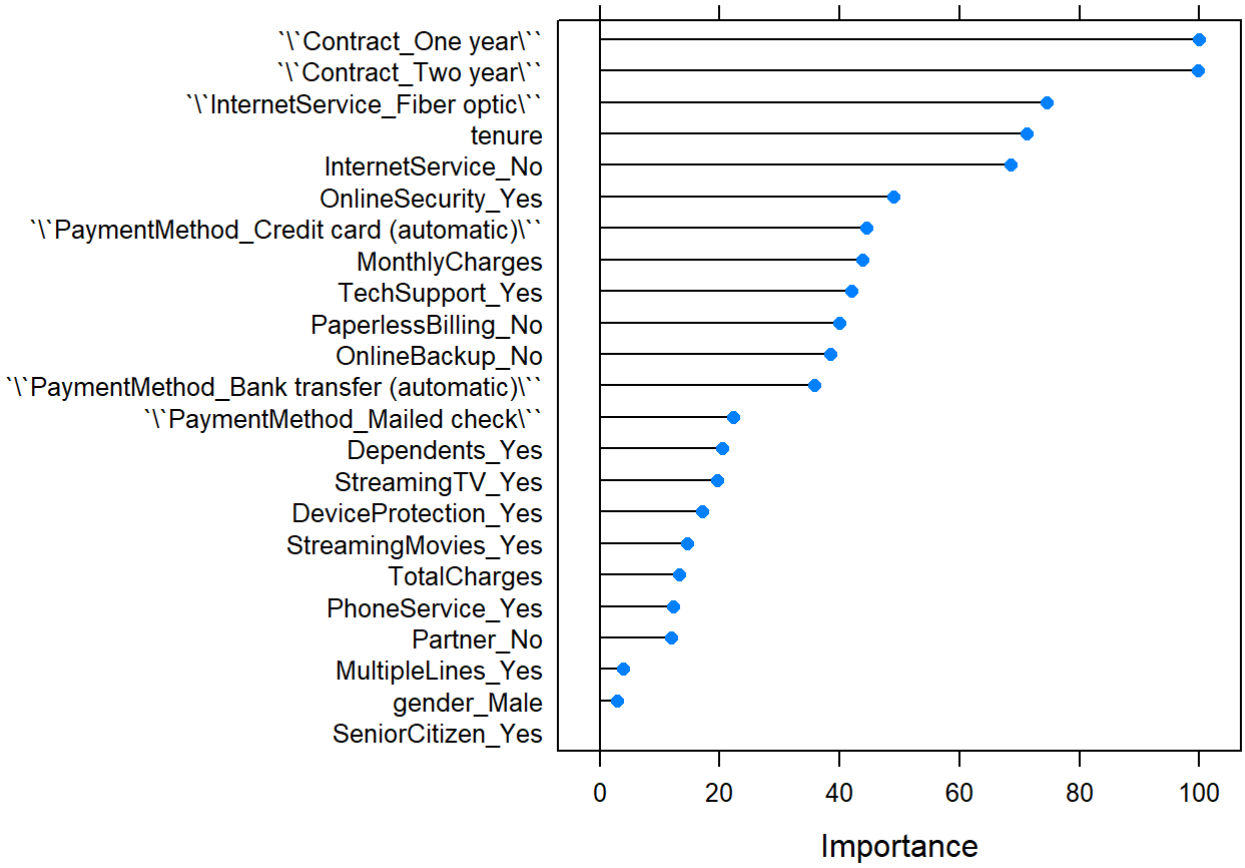
Tras todas estas pruebas, se ha mostrado la mejor solución o elección en cada caso. Una vez analizado cada algoritmo para las mejores soluciones o elecciones planteadas se ha observado que el algoritmo que mejor se adapta a la casuistica con los resultados y metricas ofrecidos es GLM.

Con este algotimo se ha logrado un auc de 76% con una sensibilidad del 71% y una especificidad del 78%. Estos datos en general no son malos, es un buen comienzo de modelado. Aún así no supera el auc y la especificidad (84% y 90%) del de H2O que era un GBM. Pero como se mencionaba antes, la sensibilidad o ser capaz de acertar el mayor numero de abandonos sin errar en exceso en los falsos negativos es bastante importante. Ya que de cara a negocio es mucho mejor detectar realmente quienes se van y fallar alguno diciendo que se va y luego no; que detectar muy bien quienes se quedan y fallar algo más diciendo que no se van cuando luego sí. Y en ese caso este algoritmo de GLM es mejor que el de H2O (71% frente a 59%) que sus metricas son altas por acertar quienes no se van.

Por tanto puede entenderse que en ciertos aspectos de eficiencia, el algoritmo de GLM es mejor que el de H2O (GBM) y por consiguiente lo supera.

Las variables más importantes de este algoritmo han sido:

```
plot(varImp(model_glm_train))
```



Como puede verse en el gráfico hay variables dummies por lo que habría que valorar tanto la variable en si como en los valores que se fija y en los ausentes. Como se puede ver la variable más importante es la de que tipo de contrato, el servicio de internert y el tiempo de estancia en la compañia. Variables que en el EDA ya se veian con una tendencia significativa. Lo que por lo tanto, viendo el gráfico se confirma las hipotesis que se lanzaban al inicio y las conclusiones plateadas tras el EDA.

Conslusiones

En la problematica de abandono de clientes para una empresa de telecomunicaciones se ha ido tratando la problematica por partes. Por un lado se ha tratado de entender quienes y porque se van a través del EDA, analizando las variables y sus indicadores. Estos han generado un perfil claro, “Los veletas inconformistas”.

Este perfil es un perfil como se ha mencionado antes, lleva poco tiempo en la compañia lo cual tiene un cargo acumulado menor, tiene un contrato que se renueva mes a mes y lo que esto implica también un cargo mensual más alto. Además es un perfil que busca satisfacer lo necesario con condiciones son bastante básicas y sin ataduras con otro cliente o dependencias, busca algo más impersonal pagando de modo electrónico.

Por todo ello se ve que son gente que va de compañia en compañia buscando un buen precio con una franja de tiempo reducida, no desea ataduras ya que pide un servicio básico como por ejemplo del de fibra óptica.

Por lo que se propondría a negocio como solución inicial estudiar la posibilidad de ofertar paquetes más básicos a un precio menor trando de captarles en el plazo de 1 año o 1 año y medio. De este modo se busca una fidelización que retornaría en ganancias más a medio largo plazo que al corto.

Todo ello teniendo en cuenta que el porcentaje a retener el de mas de un 26% de la clientela y que son muchos de ellos los que mensualmente una cuota más alta dejan en la compañia.

En cuanto al modelado para detectarlos, se ha llegado a un primer modelo final con una capacidad de detección del 76% de la cual aciertan quíen se va en un 71%. Es un buen comienzo que habría que seguir mejorando, investigando otras técnicas de feature engineering, probando con otros modelos e hiperparametros, quizás cambiando el modelo de predición de 0 o 1 a una probabilidad de ser 1 y por supuesto entrenando los algoritmos con nuevos datos.