SIM. Assignment 2: Telco Customer Churn

Adrià Casanova, Víctor Garcia, Zhengyong Ji

2024-01-05

Table of Contents

* Residual analysis at the end of the project.
* Profiling: AN EXTRA POINT TO BE DONE. WE CAN DO IT LATTER. USE THE SCRIPT PROVIDED BY MVA.

# 0. Introduction

In this project, we will study the data set “Telco Customer Churn”, which can be found at <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>. Our goal is to analyze the correlation between the amount of customers who left within the last month (Churn) and different features that describe the customer and the services he/she/they has signed up for. Then, we will build a logistic model that will allow us to predict the variable Churn.

All members have contributed equally to all parts of the project.

We start by taking a first general look at the dataset.

str(df)

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...

summary(df)

## customerID gender SeniorCitizen Partner Dependents  
## 0002-ORFBO: 1 Female:3488 Min. :0.0000 No :3641 No :4933   
## 0003-MKNFE: 1 Male :3555 1st Qu.:0.0000 Yes:3402 Yes:2110   
## 0004-TLHLJ: 1 Median :0.0000   
## 0011-IGKFF: 1 Mean :0.1621   
## 0013-EXCHZ: 1 3rd Qu.:0.0000   
## 0013-MHZWF: 1 Max. :1.0000   
## (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 Contract   
## No :2810 No :2785 Month-to-month:3875   
## No internet service:1526 No internet service:1526 One year :1473   
## Yes :2707 Yes :2732 Two year :1695   
##   
##   
##   
##   
## PaperlessBilling PaymentMethod MonthlyCharges   
## No :2872 Bank transfer (automatic):1544 Min. : 18.25   
## Yes:4171 Credit card (automatic) :1522 1st Qu.: 35.50   
## Electronic check :2365 Median : 70.35   
## Mailed check :1612 Mean : 64.76   
## 3rd Qu.: 89.85   
## Max. :118.75   
##   
## TotalCharges Churn   
## Min. : 18.8 No :5174   
## 1st Qu.: 401.4 Yes:1869   
## Median :1397.5   
## Mean :2283.3   
## 3rd Qu.:3794.7   
## Max. :8684.8   
## NA's :11

The data set contains 7043 observations of 21 variables.

# 1. Data preparation

The first part of the project consisted on doing some basic data preparation to ensure that data is ready for the next sections.

Firstly, we checked that all datatypes were consistent with the metadata and declared “SeniorCitizen” as a factor, as it represented a qualitative concept.

df$SeniorCitizen <- factor(df$SeniorCitizen, labels = c("Yes", "No"))

Secondly, we discretized all numeric variables by splitting data into 4 categories. Their boundaries were obtained simply by dividing the total range in 4 equal intervals and the distribution was checked using histograms to ensure that they were similar to the original variables.

df$c.tenure <- df$tenure # Create a new variable called Categorical.tenure  
m.tenure <- max(df$tenure, na.rm = TRUE)  
df$c.tenure <- replace(df$c.tenure, df$tenure <= m.tenure/4, m.tenure/4)  
for (i in 1:3) {  
 idx <- (m.tenure\*i/4 < df$tenure) & (df$tenure <= m.tenure\*(i+1)/4)  
 df$c.tenure <- replace(df$c.tenure, idx, m.tenure\*(i+1)/4)  
}  
min(df$tenure, na.rm = TRUE)

## [1] 0

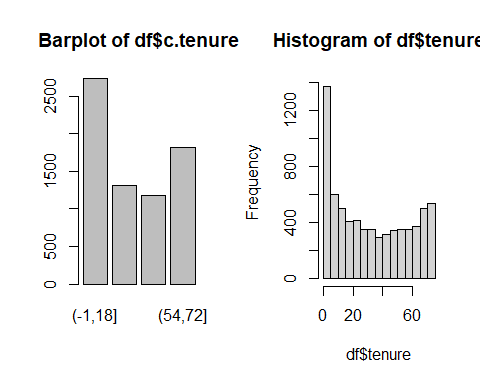
breakpts <- seq(m.tenure/4, m.tenure, m.tenure/4); breakpts

## [1] 18 36 54 72

df$c.tenure <- factor(df$c.tenure, labels = c("(-1,18]", "(18,36]",  
 "(36,54]", "(54,72]"))  
summary(df$c.tenure)

## (-1,18] (18,36] (36,54] (54,72]   
## 2734 1308 1182 1819

par(mfrow=c(1,2))  
plot(df$c.tenure, main = "Barplot of df$c.tenure")  
hist(df$tenure)



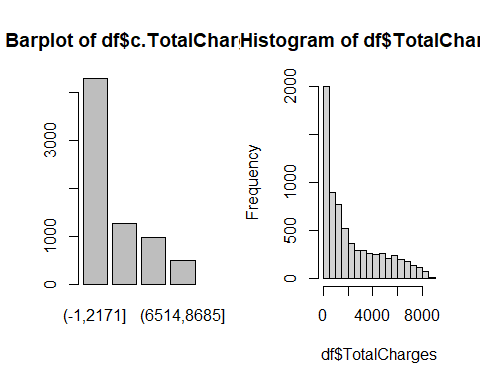
df$c.TotalCharges <- df$TotalCharges  
m.TotalCharges <- max(df$TotalCharges, na.rm = TRUE)  
df$c.TotalCharges <- replace(df$c.TotalCharges, df$TotalCharges <= m.TotalCharges/4, m.TotalCharges/4)  
for (i in 1:3) {  
 idx <- (m.TotalCharges\*i/4 < df$TotalCharges) & (df$TotalCharges <=  
 m.TotalCharges\*(i+1)/4)  
 df$c.TotalCharges <- replace(df$c.TotalCharges, idx, m.TotalCharges\*(i+1)/4)  
}  
breakpts <- seq(m.TotalCharges/4, m.TotalCharges, m.TotalCharges/4); breakpts

## [1] 2171.2 4342.4 6513.6 8684.8

df$c.TotalCharges <- factor(df$c.TotalCharges, labels = c("(-1,2171]",  
 "(2171,4342]",   
 "(4342,6514]",  
 "(6514,8685]"))  
summary(df$c.TotalCharges)

## (-1,2171] (2171,4342] (4342,6514] (6514,8685] NA's   
## 4295 1270 975 492 11

par(mfrow=c(1,2))  
plot(df$c.TotalCharges, main = "Barplot of df$c.TotalCharges")  
hist(df$TotalCharges)



df$c.MonthlyCharges <- df$MonthlyCharges  
m.MonthlyCharges <- max(df$MonthlyCharges, na.rm = TRUE)  
df$c.MonthlyCharges <- replace(df$c.MonthlyCharges, df$MonthlyCharges <= m.MonthlyCharges/4, m.MonthlyCharges/4)  
for (i in 1:3) {  
 idx <- (m.MonthlyCharges\*i/4 < df$MonthlyCharges) & (df$MonthlyCharges <=  
 m.MonthlyCharges\*(i+1)/4)  
 df$c.MonthlyCharges <- replace(df$c.MonthlyCharges, idx,  
 m.MonthlyCharges\*(i+1)/4)  
}  
min(df$MonthlyCharges, na.rm = TRUE)

## [1] 18.25

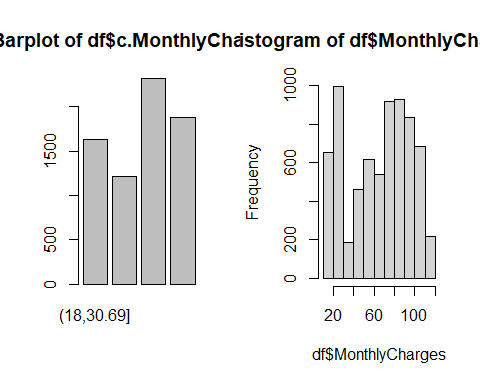
breakpts <- seq(m.MonthlyCharges/4, m.MonthlyCharges, m.MonthlyCharges/4)  
breakpts

## [1] 29.6875 59.3750 89.0625 118.7500

df$c.MonthlyCharges <- factor(df$c.MonthlyCharges, labels = c("(18,30.69]",  
 "(30.69,59.38]",   
 "(59.38,89.06]",  
 "(89.06,118.75]"))  
summary(df$c.MonthlyCharges)

## (18,30.69] (30.69,59.38] (59.38,89.06] (89.06,118.75]   
## 1634 1208 2317 1884

par(mfrow=c(1,2))  
plot(df$c.MonthlyCharges, main = "Barplot of df$c.MonthlyCharges")  
hist(df$MonthlyCharges)



par(mfrow=c(1,1))

Lastly, we identified categorical and numerical variables for later use.

numeric\_val\_idx = which(sapply(df, is.numeric))  
numeric\_val = names(df)[numeric\_val\_idx]  
# The only numerical features that we have are tenure, MonthlyCharges and TotalChages.  
  
# So the remaining will be categorical features.  
categoric\_val\_idx = which(sapply(df, is.factor))  
categoric\_val = names(df)[categoric\_val\_idx]

# 2. Exploratory Data Analysis (EDA)

EDA was done mainly automatically using the “DataExplorer” library. It plots, for each variable, the distribution of numeric variables, the proportion of individuals in each category and the amount of missing values, among other metadata.

The main conclusions of this section are: 1- Using the QQ plots and distribution plots we see that no numerical variable is normally distributed. This was also checked visually and with Kolmogorov-Smirnov tests, a more suitable approach than Shappiro-Wilk for large samples.

2- Our database is not balanced in some categories, like PhoneService (10% of “No”) or SeniorCitizen(16% of “No”). This is specially relevant for the target, “Churn”, that has 73% of cases of “No”, so individuals that churned will be more difficult to predict.

3- Qualitative variables have a maximum of 4 levels, so all of them may be suitable for modeling without any aggregation.

5- Some categories, like “OnlineSecurity” or “OnlineBackup”, are not applicable if the client does not have an internet connection. Consequently, there is a special level for those cases that contains around 22% of the clients.

# Basic EDA  
summary(df)

## customerID gender SeniorCitizen Partner Dependents  
## 0002-ORFBO: 1 Female:3488 Yes:5901 No :3641 No :4933   
## 0003-MKNFE: 1 Male :3555 No :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 Contract   
## No :2810 No :2785 Month-to-month:3875   
## No internet service:1526 No internet service:1526 One year :1473   
## Yes :2707 Yes :2732 Two year :1695   
##   
##   
##   
##   
## PaperlessBilling PaymentMethod MonthlyCharges   
## No :2872 Bank transfer (automatic):1544 Min. : 18.25   
## Yes:4171 Credit card (automatic) :1522 1st Qu.: 35.50   
## Electronic check :2365 Median : 70.35   
## Mailed check :1612 Mean : 64.76   
## 3rd Qu.: 89.85   
## Max. :118.75   
##   
## TotalCharges Churn c.tenure c.TotalCharges  
## Min. : 18.8 No :5174 (-1,18]:2734 (-1,2171] :4295   
## 1st Qu.: 401.4 Yes:1869 (18,36]:1308 (2171,4342]:1270   
## Median :1397.5 (36,54]:1182 (4342,6514]: 975   
## Mean :2283.3 (54,72]:1819 (6514,8685]: 492   
## 3rd Qu.:3794.7 NA's : 11   
## Max. :8684.8   
## NA's :11   
## c.MonthlyCharges  
## (18,30.69] :1634   
## (30.69,59.38] :1208   
## (59.38,89.06] :2317   
## (89.06,118.75]:1884   
##   
##   
##

# Completed EDA   
#create\_report(df, output\_file = "Telco.html")

# tests  
ks.test(df$TotalCharges, "pnorm")

## Warning in ks.test.default(df$TotalCharges, "pnorm"): ties should not be  
## present for the Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: df$TotalCharges  
## D = 1, p-value < 2.2e-16  
## alternative hypothesis: two-sided

ks.test(df$MonthlyCharges, "pnorm")

## Warning in ks.test.default(df$MonthlyCharges, "pnorm"): ties should not be  
## present for the Kolmogorov-Smirnov test

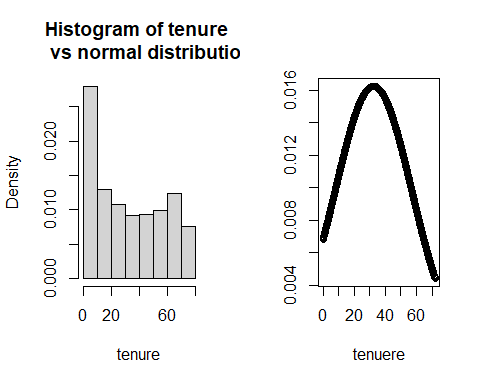
##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: df$MonthlyCharges  
## D = 1, p-value < 2.2e-16  
## alternative hypothesis: two-sided

ks.test(df$tenure, "pnorm")

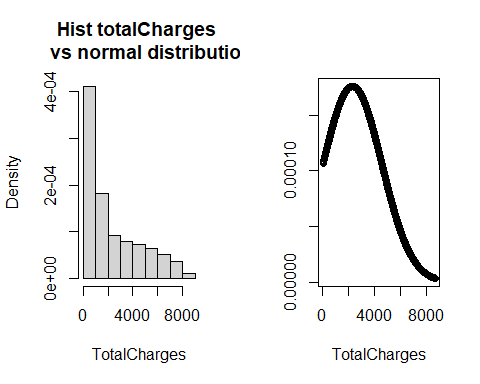
## Warning in ks.test.default(df$tenure, "pnorm"): ties should not be present for  
## the Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: df$tenure  
## D = 0.88865, p-value < 2.2e-16  
## alternative hypothesis: two-sided

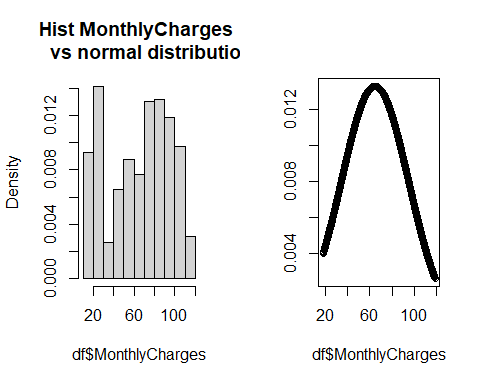
# plots  
par(mfrow=c(1,2))  
hist(df$tenure, prob = TRUE, breaks = 10, main = 'Histogram of tenure   
 vs normal distribution', xlab = 'tenure')  
x <- seq(min(df$tenure), max(df$tenure), by = .1)  
y <- dnorm(x, mean = mean(df$tenure), sd = sd(df$tenure))  
plot(x,y, xlab = 'tenuere', ylab = '')



hist(df$TotalCharges, prob = TRUE, breaks = 10, main = 'Hist totalCharges   
 vs normal distribution', xlab = 'TotalCharges')  
x <- seq(min(df$TotalCharges, na.rm = TRUE), max(df$TotalCharges, na.rm = TRUE),  
 by = 10)  
y <- dnorm(x, mean = mean(df$TotalCharges, na.rm = TRUE), sd = sd(df$TotalCharges, na.rm = TRUE))  
plot(x,y, xlab = 'TotalCharges', ylab = '')



hist(df$MonthlyCharges, prob = TRUE, breaks = 10, main = 'Hist MonthlyCharges   
 vs normal distribution', xlab = 'df$MonthlyCharges')  
x <- seq(min(df$MonthlyCharges, na.rm = TRUE), max(df$MonthlyCharges, na.rm = TRUE),  
 by = .1)  
y <- dnorm(x, mean = mean(df$MonthlyCharges, na.rm = TRUE), sd = sd(df$MonthlyCharges, na.rm = TRUE))  
plot(x,y, xlab = 'df$MonthlyCharges', ylab = '')



par(mfrow=c(1,1))

# 3. Data Quality Report

In this section we analysed the missing values, outliers and errors of numeric variables to increase the quality of data before modeling.

To start with, we detected that only “TotalCharges”, and hence “c.TotalCharges”, has a total of 22 missing observations. However, all of them correspond to new clients who have not receive their first invoice yet, so “TotalCharges” can not have a value. In other words, they are “not applicable cases”. We naturally impute this observations with 0.

# Distribution of missings in df per variable  
apply(sapply(df, is.na), 2, sum)

## customerID gender SeniorCitizen Partner   
## 0 0 0 0   
## Dependents tenure PhoneService MultipleLines   
## 0 0 0 0   
## InternetService OnlineSecurity OnlineBackup DeviceProtection   
## 0 0 0 0   
## TechSupport StreamingTV StreamingMovies Contract   
## 0 0 0 0   
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges   
## 0 0 0 11   
## Churn c.tenure c.TotalCharges c.MonthlyCharges   
## 0 0 11 0

# Distribution of missings in df per individual  
table(apply(sapply(df, is.na), 1, sum))

##   
## 0 2   
## 7032 11

# Check that all missings in "TotalCharges" correspond to individuals tenure = 0  
TotalCharges.na <- which(is.na(df$TotalCharges))  
sum(TotalCharges.na == which(df$tenure == 0)) == length(TotalCharges.na)

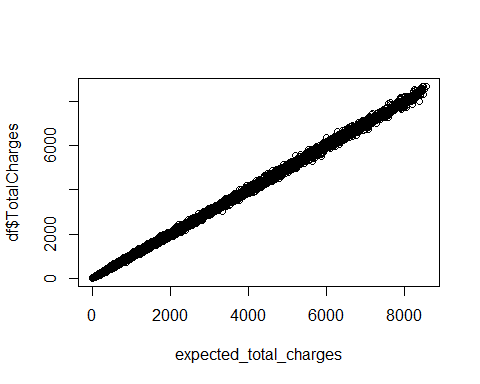
## [1] TRUE

# So we transform them after creating a new numeric variable with all the missings of the database  
df$n.na <- apply(sapply(df, is.na), 1, sum)  
  
df$TotalCharges[TotalCharges.na] = 0  
df$c.TotalCharges[TotalCharges.na] = "(-1,2171]"

Secondly, we detected data inconsistencies. For categorical values, we checked the EDA automatic reports and the summaries to ensure that all qualitative variables categories were meaningful and that there was not any misspelling errors. We also checked that all values of numeric variables were positive and reasonable.

Additionally, for “TotalCharges” we ensured that all the values were correct by manually calculating the value and comparing it to the actual total charge.

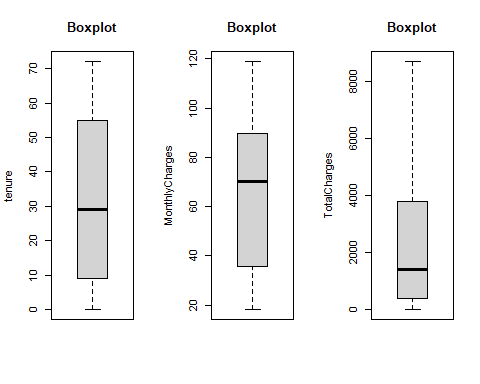
# Expected total charges as the product of monthly charges and tenure  
expected\_total\_charges = df$MonthlyCharges \* df$tenure  
  
# Plot them against the actual total charges  
plot(expected\_total\_charges, df$TotalCharges)



# There are no outliers, so TotalCharges is consistent.

Thirdly, we analysed univariate outliers in numeric variables using Boxplots and the typical thresholds: 1.5 \* IQR(interquartile range) for mild outliers and 3 \* IQR for severe outliers. As there were not any we considered that all points were suitable for our models.

par(mfrow=c(1, length(numeric\_val\_idx)))  
for (var in as.numeric(numeric\_val\_idx)) {  
 Boxplot(df[,var], ylab = names(df)[var], main = "Boxplot")   
}



par(mfrow=c(1,1))

## 3.1 In depth analysis of missing values

Next, we will compute for every group of individuals the mean of missing values. Then we will rank the groups according to the computed mean.

# c.TotalCharges has missings, so it doesn't make sense to compute the mean  
# of missings in its categories  
  
interesting\_cat\_idx <- categoric\_val\_idx[-c(1,20)]  
k = 0  
for (i in interesting\_cat\_idx){  
 k <- k + length(levels(df[,i]))  
}  
groups.na <- matrix(0, k, 2)  
l = 1  
for (idx in interesting\_cat\_idx) {  
 categories.na <- tapply(df$n.na, df[,idx], mean)  
 for (j in seq(length(categories.na))) {  
 groups.na[l + j - 1,] <- c(categories.na[j],  
 paste(names(df)[idx], levels(df[,idx])[j],  
 sep = "."))  
 }  
 l <- l + j  
}  
groups.na.df <- data.frame(na.perc = groups.na[,1], group = groups.na[,2])  
groups.na.df[order(groups.na.df$na.perc, decreasing = TRUE),]

## na.perc group  
## 37 0.0117994100294985 Contract.Two year  
## 8 0.0104265402843602 Dependents.Yes  
## 43 0.00992555831265509 PaymentMethod.Mailed check  
## 46 0.00804681784930505 c.tenure.(-1,18]  
## 16 0.00786369593709043 InternetService.No  
## 18 0.00786369593709043 OnlineSecurity.No internet service  
## 21 0.00786369593709043 OnlineBackup.No internet service  
## 24 0.00786369593709043 DeviceProtection.No internet service  
## 27 0.00786369593709043 TechSupport.No internet service  
## 30 0.00786369593709043 StreamingTV.No internet service  
## 33 0.00786369593709043 StreamingMovies.No internet service  
## 50 0.00734394124847001 c.MonthlyCharges.(18,30.69]  
## 9 0.00586510263929619 PhoneService.No  
## 12 0.00586510263929619 MultipleLines.No phone service  
## 38 0.00557103064066852 PaperlessBilling.No  
## 6 0.00529100529100529 Partner.Yes  
## 44 0.00425202937765752 Churn.No  
## 14 0.00413052457662123 InternetService.DSL  
## 19 0.00396235760277365 OnlineSecurity.Yes  
## 28 0.00391389432485323 TechSupport.Yes  
## 3 0.00372818166412472 SeniorCitizen.Yes  
## 2 0.00337552742616034 gender.Male  
## 51 0.0033112582781457 c.MonthlyCharges.(30.69,59.38]  
## 25 0.00330305532617671 DeviceProtection.Yes  
## 22 0.00329353643474681 OnlineBackup.Yes  
## 31 0.00295530107129664 StreamingTV.Yes  
## 11 0.00294985250737463 MultipleLines.No  
## 32 0.00287253141831239 StreamingMovies.No  
## 1 0.00286697247706422 gender.Female  
## 10 0.00282974375098255 PhoneService.Yes  
## 13 0.00269269606193201 MultipleLines.Yes  
## 40 0.00259067357512953 PaymentMethod.Bank transfer (automatic)  
## 52 0.00258955545964609 c.MonthlyCharges.(59.38,89.06]  
## 39 0.00143850395588588 PaperlessBilling.Yes  
## 36 0.00135777325186694 Contract.One year  
## 41 0.00131406044678055 PaymentMethod.Credit card (automatic)  
## 5 0.00109859928591046 Partner.No  
## 34 0.000732064421669107 StreamingMovies.Yes  
## 29 0.000711743772241993 StreamingTV.No  
## 20 0.000647668393782383 OnlineBackup.No  
## 23 0.000646203554119548 DeviceProtection.No  
## 26 0.000575871004894904 TechSupport.No  
## 17 0.000571755288736421 OnlineSecurity.No  
## 4 0 SeniorCitizen.No  
## 7 0 Dependents.No  
## 15 0 InternetService.Fiber optic  
## 35 0 Contract.Month-to-month  
## 42 0 PaymentMethod.Electronic check  
## 45 0 Churn.Yes  
## 47 0 c.tenure.(18,36]  
## 48 0 c.tenure.(36,54]  
## 49 0 c.tenure.(54,72]  
## 53 0 c.MonthlyCharges.(89.06,118.75]

The groups with the highest proportion of missing data are made of those individuals who:

* Have a two-year contract
* Have dependents
* Pay with a mailed check

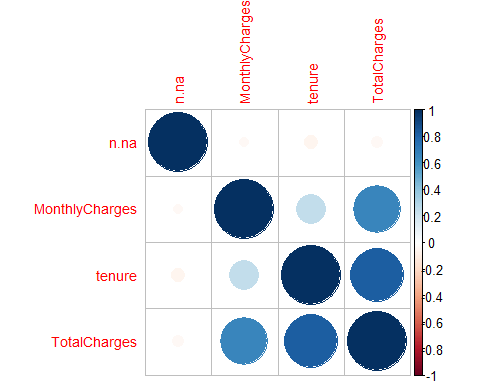
Since the set of individuals with missing data is exactly that of the new clients, we conclude that recently incorporated clients tend to: sign a two-year contract, have dependents and pay with a mailed check.

We can compute as well the pearson correlation coefficient between “n.na” and the numerical variables.

# Creation of the correlation matrix  
corr\_mat <- cor(df[,c(numeric\_val\_idx, 25)],)  
corr\_mat

## tenure MonthlyCharges TotalCharges n.na  
## tenure 1.00000000 0.24789986 0.82617840 -0.05213467  
## MonthlyCharges 0.24789986 1.00000000 0.65117383 -0.03068535  
## TotalCharges 0.82617840 0.65117383 1.00000000 -0.03977955  
## n.na -0.05213467 -0.03068535 -0.03977955 1.00000000

corrplot(corr\_mat, order = 'hclust', tl.cex = 0.9)

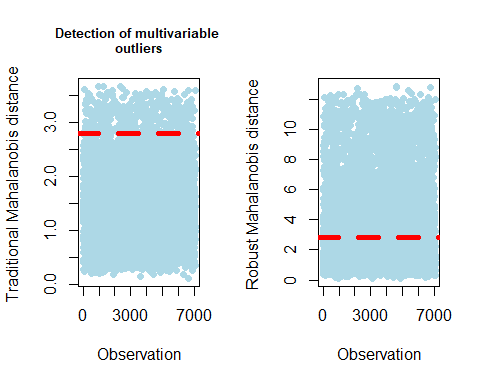


n.na is independent to the rest of numerical variables, probably because it evaluates to 0 in most observations.

## 3.2 Multivariate outliers

In this section we focused on detecting the multivariate outliers using “Moutlier”. We discovered 344 multivariate outliers, about 5% of the individuals, as it was expected. We decided to maintain them and only remove them in the modeling step if they turned out to be influential points.

set.seed(123)  
res.mout <- Moutlier(df[,numeric\_val\_idx], quantile = 0.95, plot= FALSE)  
  
# Visual representation  
par(mfrow=c(1,2), cex.main=0.8)  
plot(res.mout$md, col="lightblue", pch = 19, main = 'Detection of multivariable   
outliers', xlab= 'Observation',   
 ylab ='Traditional Mahalanobis distance ')  
abline(h = res.mout$cutoff, col = "red", lwd = 5, lty = 2)  
  
plot(res.mout$rd, col="lightblue", pch = 19, xlab= 'Observation',   
 ylab ='Robust Mahalanobis distance ')  
abline(h = res.mout$cutoff, col = "red", lwd = 5, lty = 2)



par(mfrow=c(1,1), cex.main=1)  
  
# Identification of the outliers  
outliers = which(res.mout$md>res.mout$cutoff & res.mout$rd > res.mout$cutoff)   
length(outliers)

## [1] 344

length(outliers)/dim(df)[1]\*100

## [1] 4.884282

# 5. Profiling of the target and feature selection

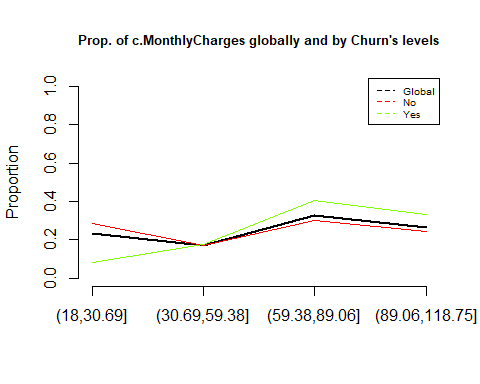
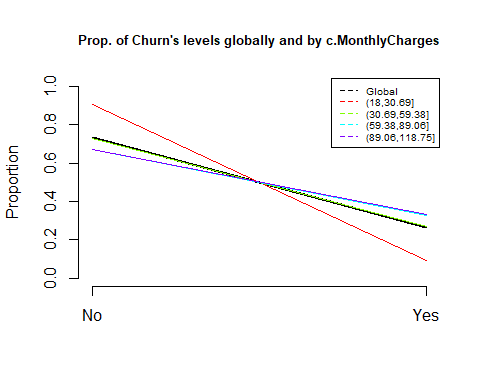
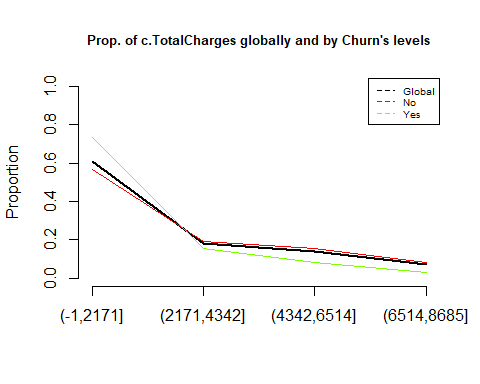
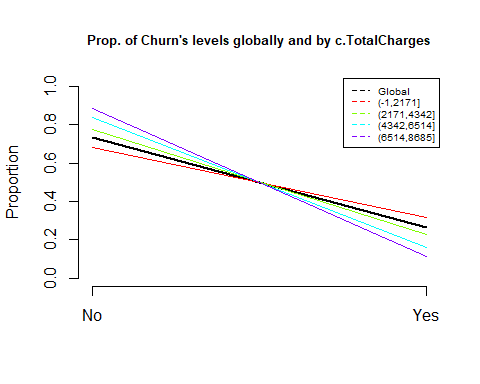
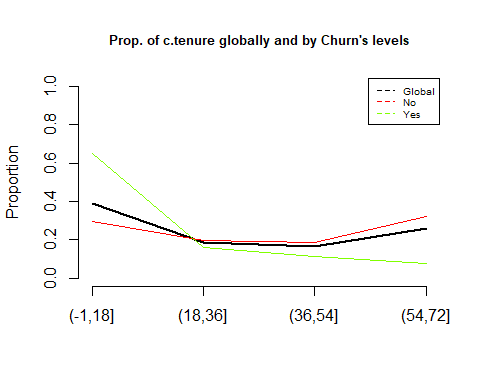
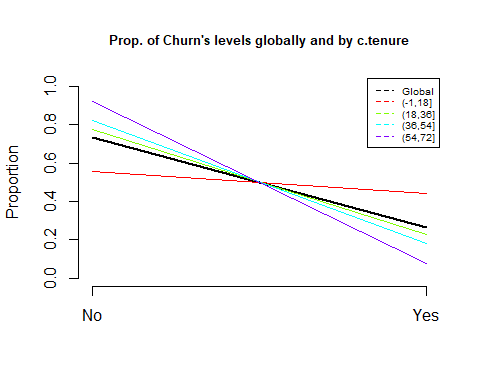
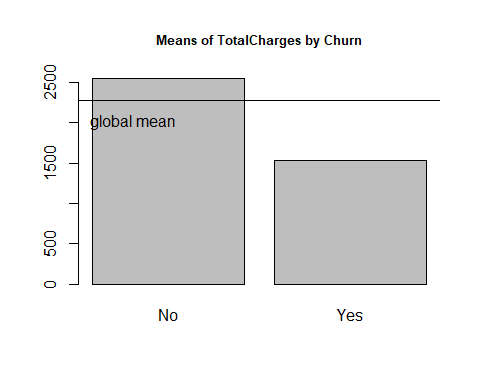
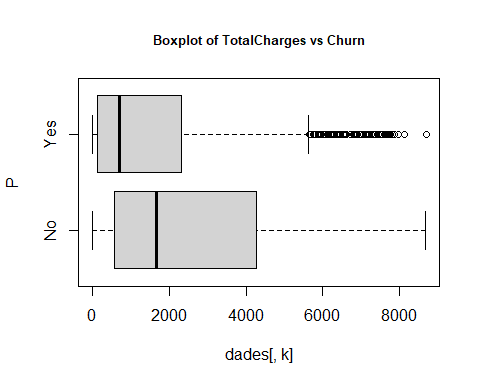
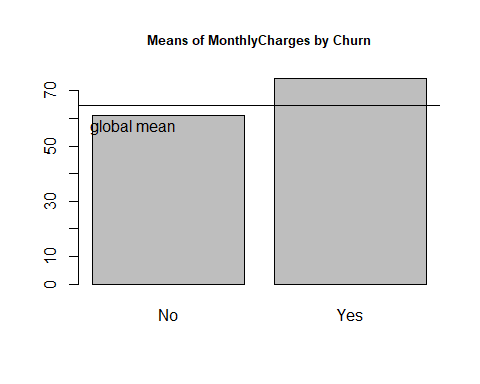
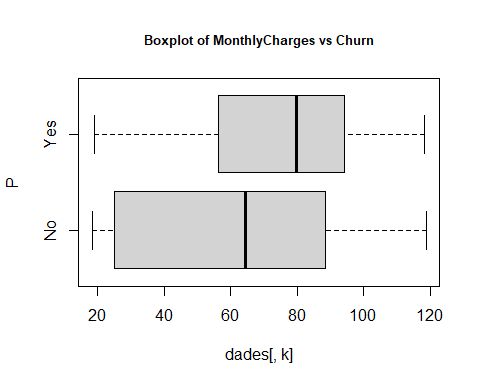
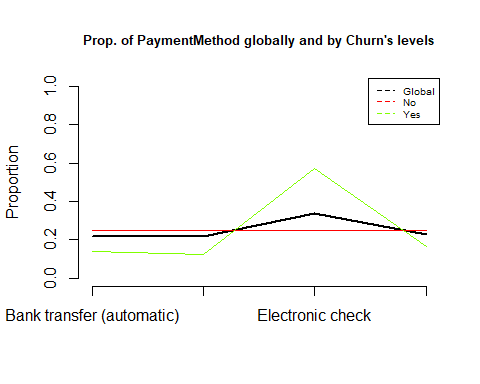
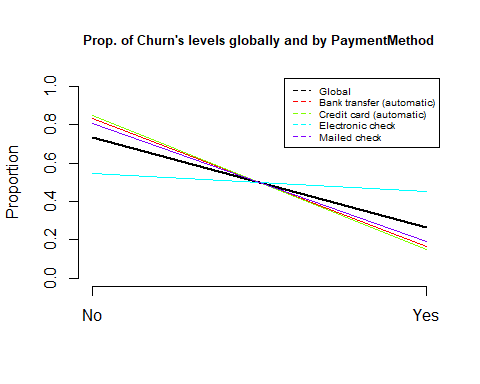
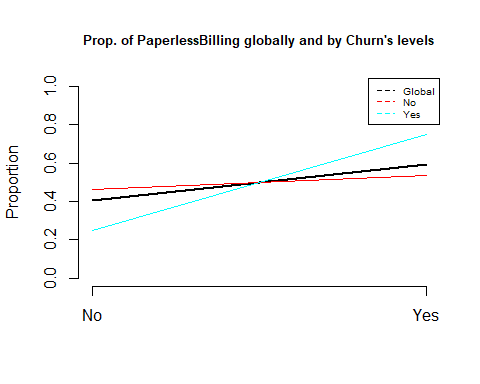
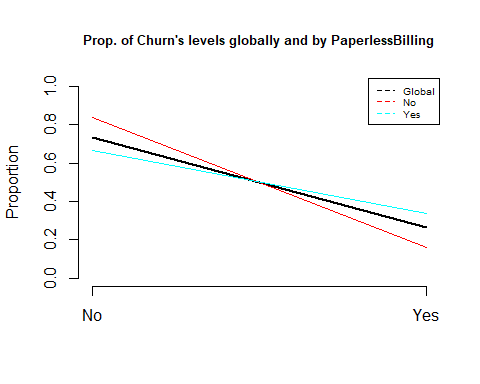
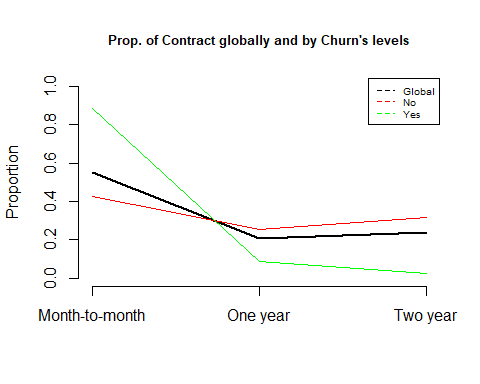
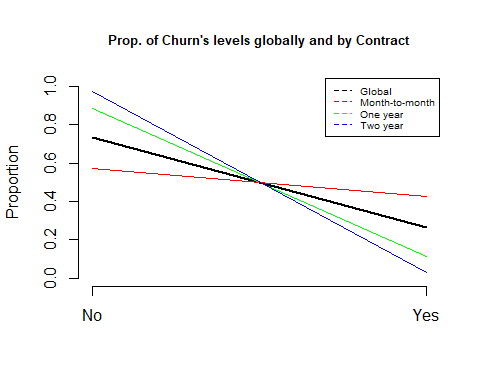
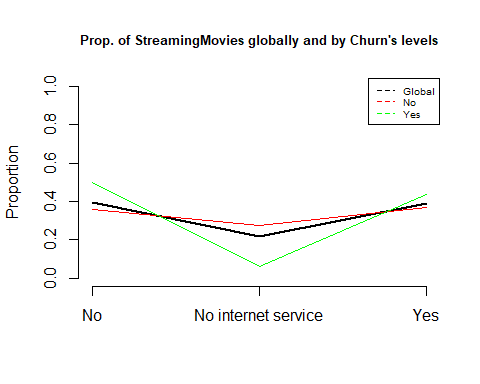
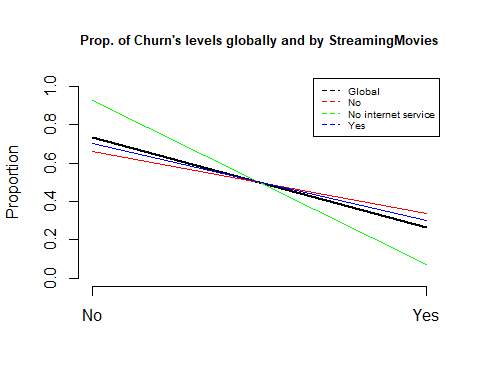
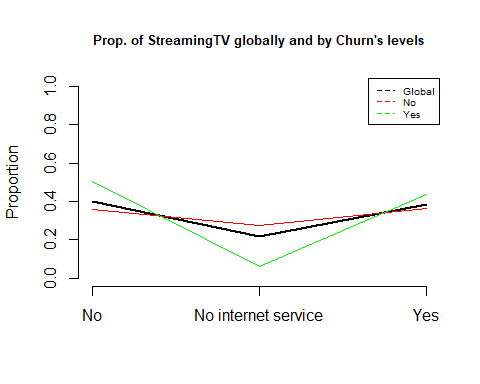
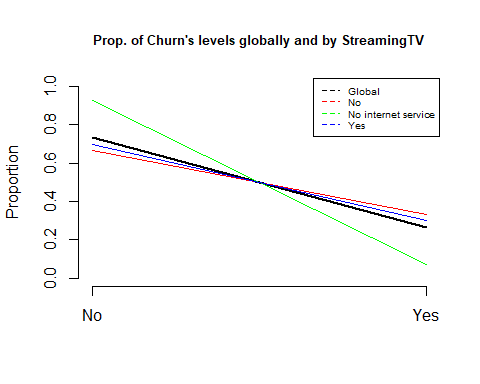
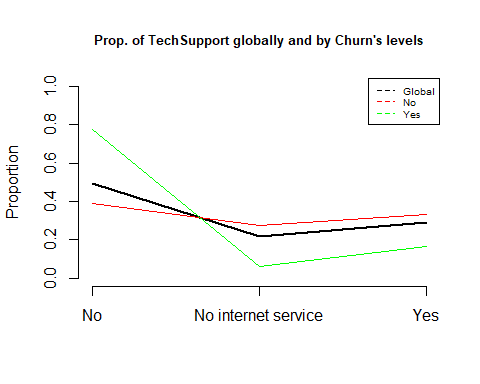
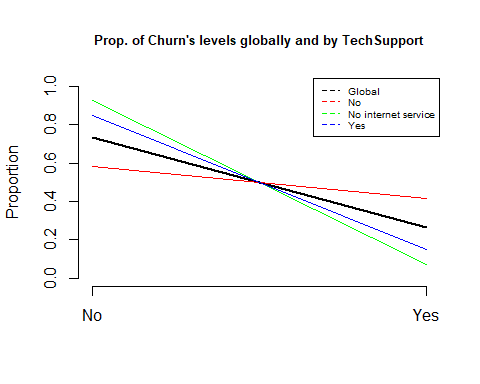
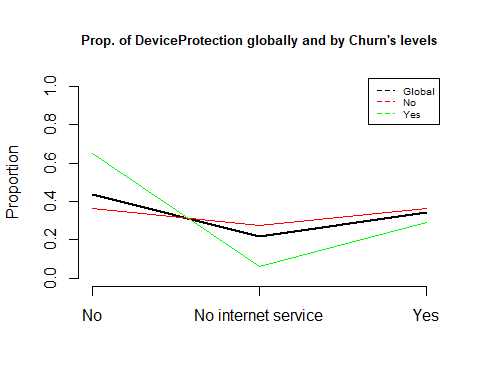
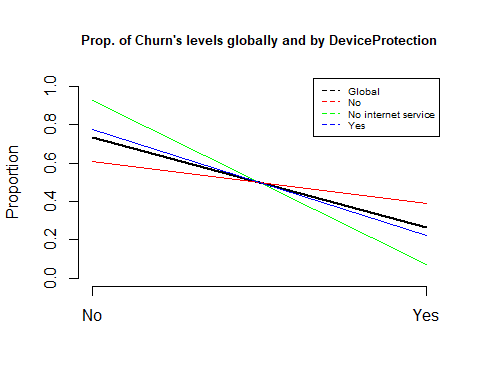
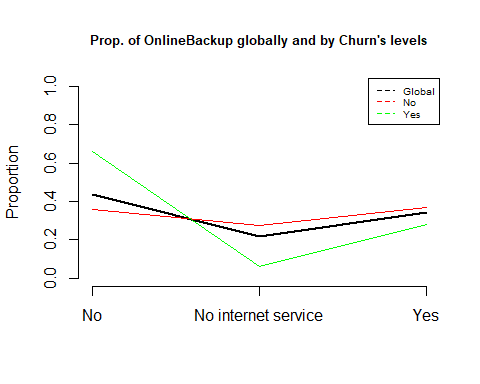
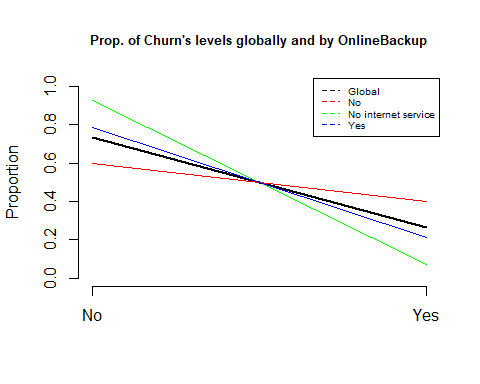
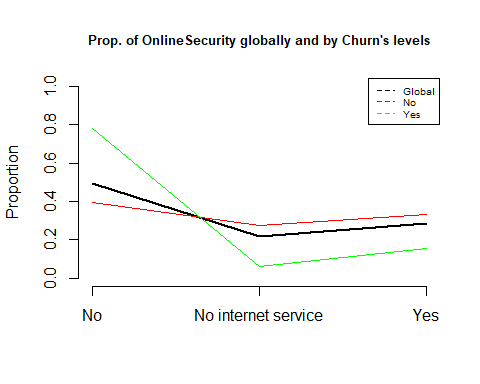
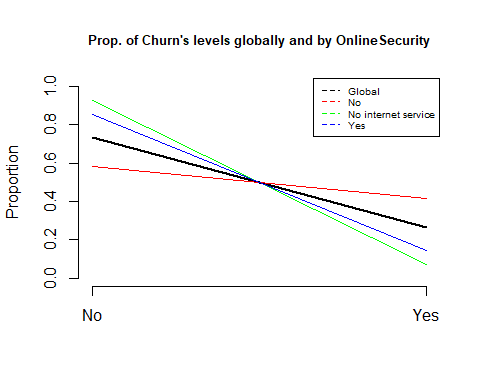
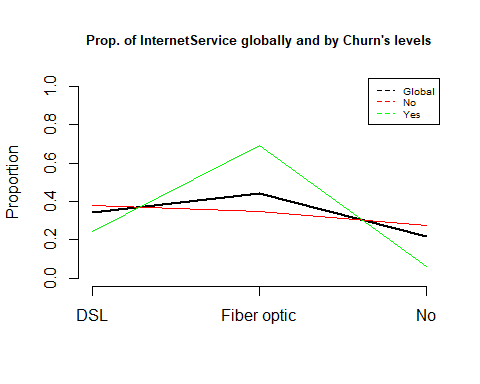
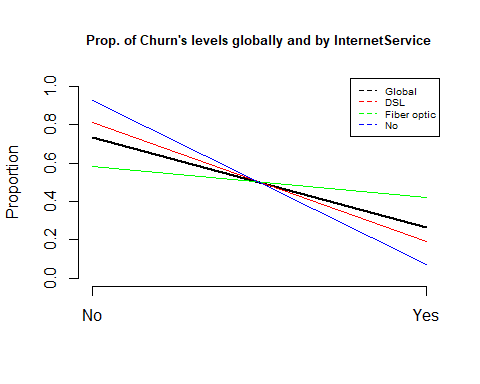
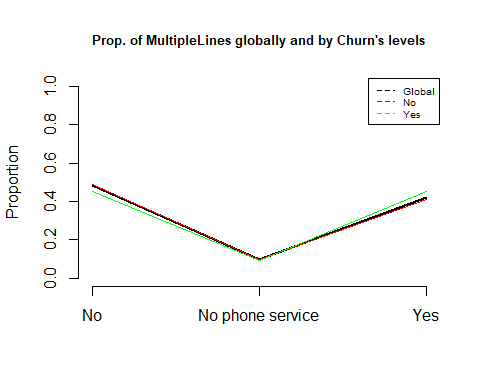
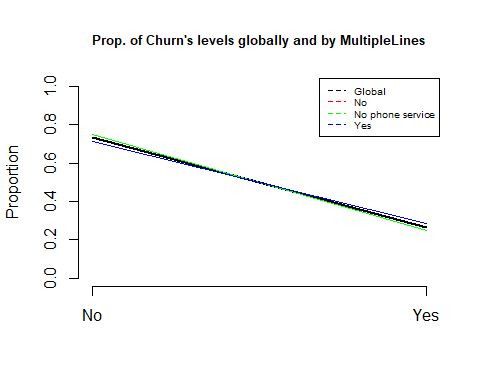
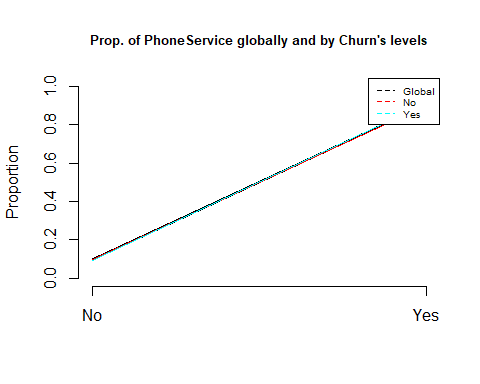
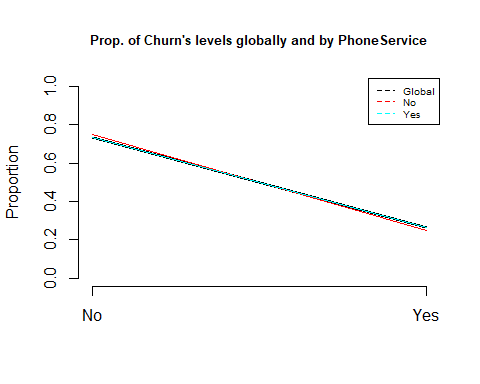
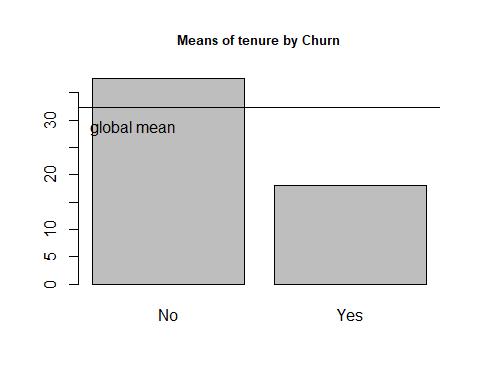
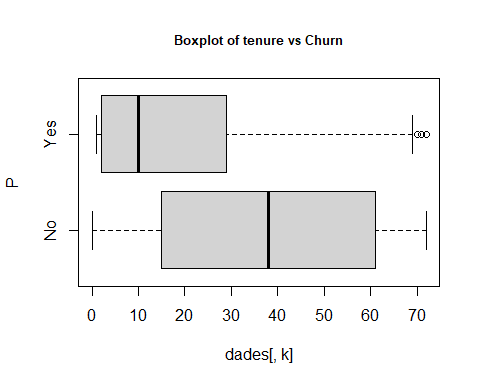
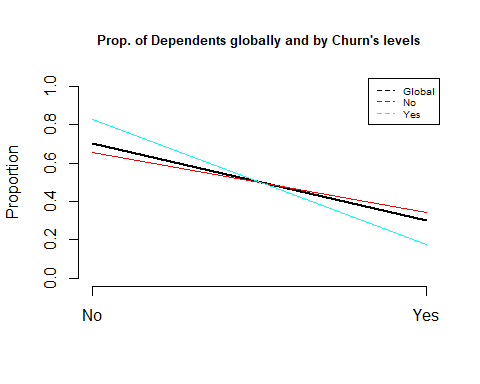
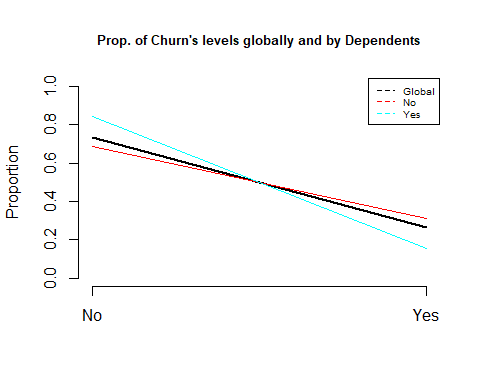
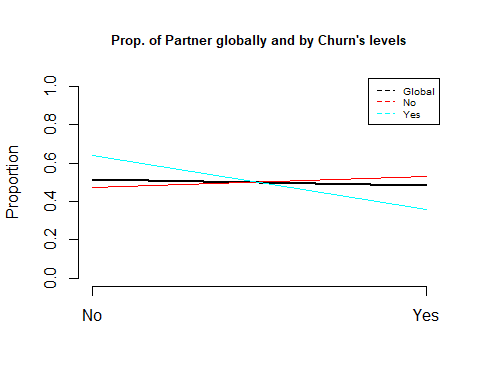
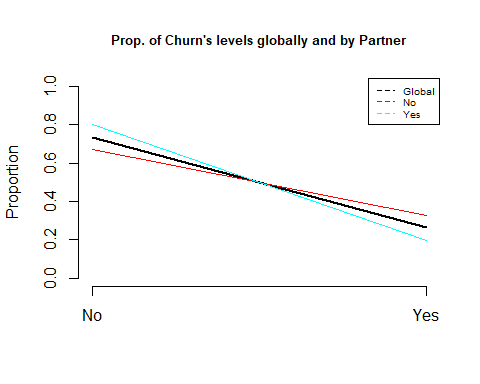
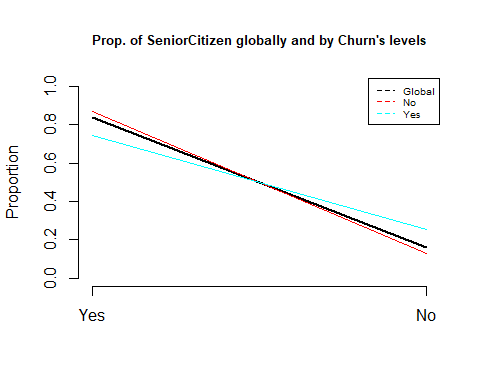
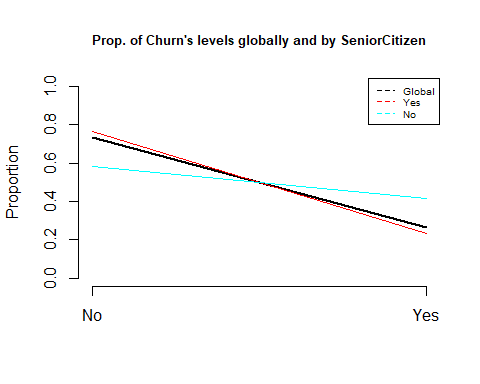
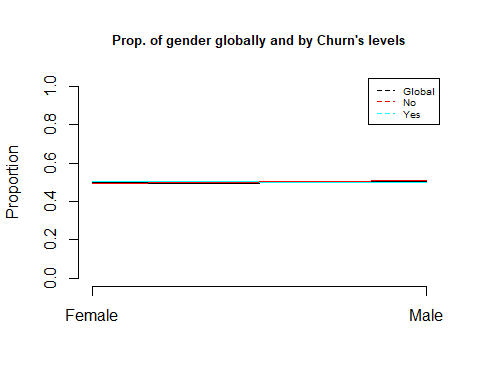
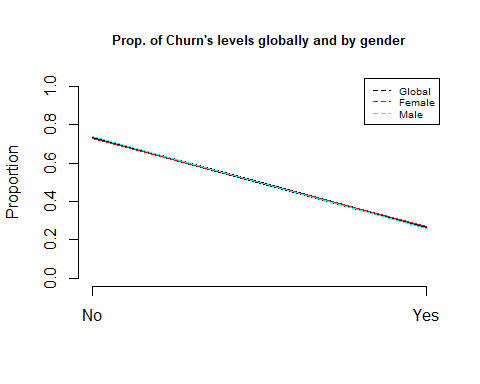
## Numeric variables’ correlations

We analysed the pearson correlation coefficient to detect variables that were highly related and not include them all in the model. In the correlation plot of section 3.1 we see that “TotalCharges” is highly correlated with “MonthlyCharges” and “tenure” as the first one is calculated as the product of the others.

## Profiling of the target

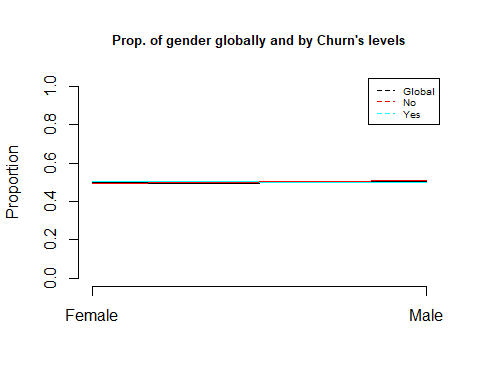
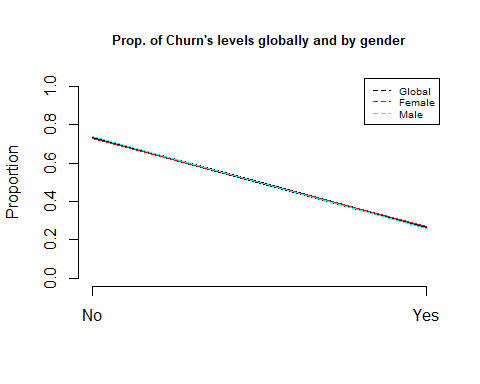
Later on, we profiled the target Churn using a custom function “profiling()” created in the Multivariate Analysis subject of the Master’s degree. This method expands “catdes()” and performs many plots and tests according to the type of each variable. We will focus on plots and the given tests’ results: Chi^2, ANOVA and Kruskal-Wallis.

# Analysis of all variables except the ID  
profiling(df[-c(1, grep("Churn", names(df)))], df$Churn, "Churn")

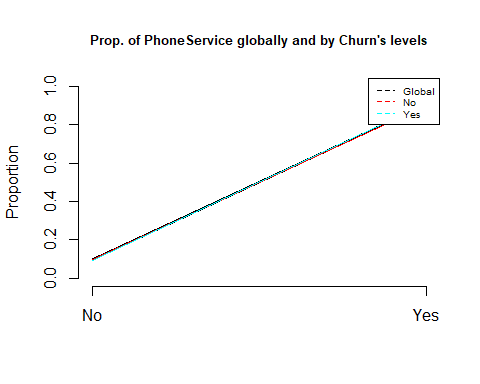
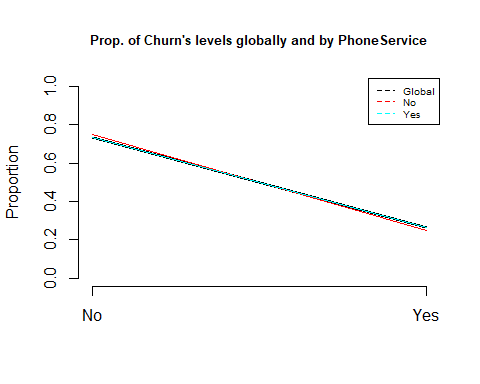


The most relevant conclusions are: - Some variables are not significant, like Gender (Chi^2 p-value=0.4866) or Phone service (Chi^2 p-value=0.3388). Consequently, we state that churn is independent of the client’s gender and whether he/she/they has a phone service contracted.

## [1] "Variable gender"  
## [1] "Categories=" "Female" "Male"



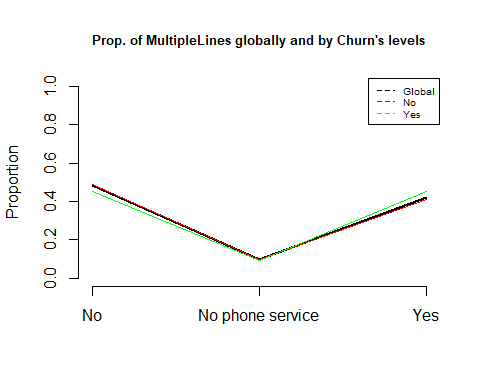
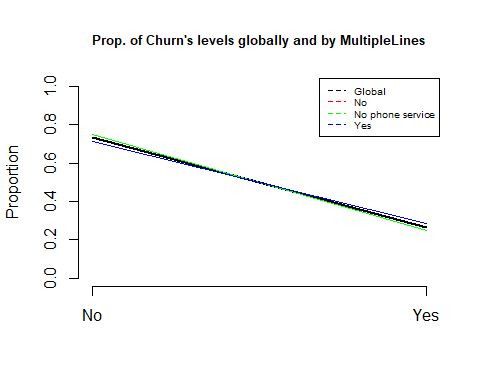
## [1] "Cross Table:"  
## P  
## No Yes  
## Female 2549 939  
## Male 2625 930  
## [1] "Distributions by columns:"  
##   
## P Female Male  
## No 0.7307913 0.7383966  
## Yes 0.2692087 0.2616034  
## [1] "Chi^2 test: "  
##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: dades[, k] and as.factor(P)  
## X-squared = 0.48408, df = 1, p-value = 0.4866  
##   
## [1] "ValorTestXquali:"  
## $rowpf  
## Xquali  
## P Female Male  
## No 0.4926556 0.5073444  
## Yes 0.5024077 0.4975923  
##   
## $vtest  
## Xquali  
## P Female Male  
## No -0.7227493 0.7227493  
## Yes 0.7227493 -0.7227493  
##   
## $pval  
## Xquali  
## P Female Male  
## No 0.234917 0.234917  
## Yes 0.234917 0.234917  
##   
## [1] "Variable PhoneService"  
## [1] "Categories=" "No" "Yes"



## [1] "Cross Table:"  
## P  
## No Yes  
## No 512 170  
## Yes 4662 1699  
## [1] "Distributions by columns:"  
##   
## P No Yes  
## No 0.7507331 0.7329036  
## Yes 0.2492669 0.2670964  
## [1] "Chi^2 test: "  
##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: dades[, k] and as.factor(P)  
## X-squared = 0.91503, df = 1, p-value = 0.3388  
##   
## [1] "ValorTestXquali:"  
## $rowpf  
## Xquali  
## P No Yes  
## No 0.09895632 0.90104368  
## Yes 0.09095773 0.90904227  
##   
## $vtest  
## Xquali  
## P No Yes  
## No 1.002202 -1.002202  
## Yes -1.002202 1.002202  
##   
## $pval  
## Xquali  
## P No Yes  
## No 0.1581231 0.1581231  
## Yes 0.1581231 0.1581231  
##   
## [1] "P.values per class: No"  
## gender PhoneService   
## 0 0   
## [1] "P.values per class: Yes"  
## gender PhoneService   
## 0 0

* There are variables like “MultipleLines” that even being significant (Chi^2 p-value=0.003464) the difference among levels is small, as we can see in the plots

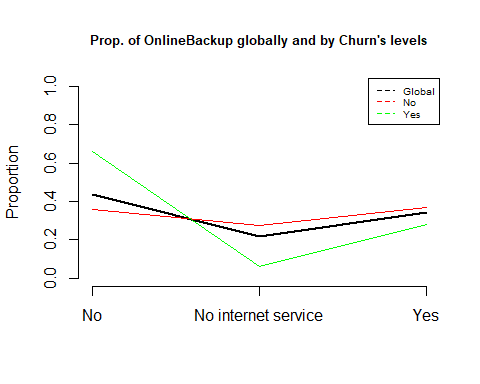
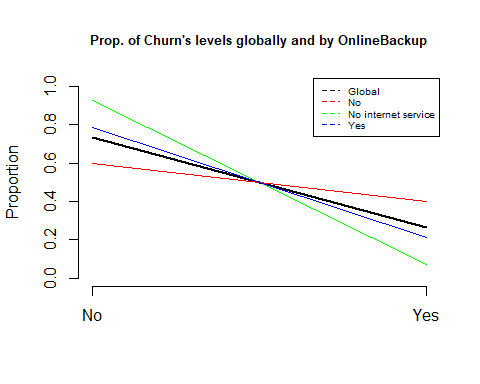
## [1] "Variable MultipleLines"  
## [1] "Categories=" "No" "No phone service" "Yes"



## [1] "Cross Table:"  
## P  
## No Yes  
## No 2541 849  
## No phone service 512 170  
## Yes 2121 850  
## [1] "Distributions by columns:"  
##   
## P No No phone service Yes  
## No 0.7495575 0.7507331 0.7139010  
## Yes 0.2504425 0.2492669 0.2860990  
## [1] "Chi^2 test: "  
##   
## Pearson's Chi-squared test  
##   
## data: dades[, k] and as.factor(P)  
## X-squared = 11.33, df = 2, p-value = 0.003464  
##   
## [1] "ValorTestXquali:"  
## $rowpf  
## Xquali  
## P No No phone service Yes  
## No 0.49110939 0.09895632 0.40993429  
## Yes 0.45425361 0.09095773 0.45478866  
##   
## $vtest  
## Xquali  
## P No No phone service Yes  
## No 2.733239 1.002202 -3.365474  
## Yes -2.733239 -1.002202 3.365474  
##   
## $pval  
## Xquali  
## P No No phone service Yes  
## No 0.0031357380 0.1581230658 0.0003820611  
## Yes 0.0031357380 0.1581230658 0.0003820611  
##   
## [1] "P.values per class: No"  
## [1] 0  
## [1] "P.values per class: Yes"  
## [1] 0

* The rest of variables, including the discretized, have a small p-value (< 2.2e-16) in the Chi^2, ANOVA or Kruskal-Wallis test, according to their type, and have at least one level where the target’s distribution is different than in the rest. For example, 40% of people that did not have an online backup churned, while only 22% of customers having the backup did.

## [1] "Variable OnlineBackup"  
## [1] "Categories=" "No" "No internet service"  
## [4] "Yes"



## [1] "Cross Table:"  
## P  
## No Yes  
## No 1855 1233  
## No internet service 1413 113  
## Yes 1906 523  
## [1] "Distributions by columns:"  
##   
## P No No internet service Yes  
## No 0.6007124 0.9259502 0.7846851  
## Yes 0.3992876 0.0740498 0.2153149  
## [1] "Chi^2 test: "  
##   
## Pearson's Chi-squared test  
##   
## data: dades[, k] and as.factor(P)  
## X-squared = 601.81, df = 2, p-value < 2.2e-16  
##   
## [1] "ValorTestXquali:"  
## $rowpf  
## Xquali  
## P No No internet service Yes  
## No 0.35852339 0.27309625 0.36838036  
## Yes 0.65971108 0.06046014 0.27982879  
##   
## $vtest  
## Xquali  
## P No No internet service Yes  
## No -22.491687 19.125155 6.903041  
## Yes 22.491687 -19.125155 -6.903041  
##   
## $pval  
## Xquali  
## P No No internet service Yes  
## No 0.000000e+00 7.795425e-82 2.545045e-12  
## Yes 2.502984e-112 0.000000e+00 2.545075e-12  
##   
## [1] "P.values per class: No"  
## [1] 0  
## [1] "P.values per class: Yes"  
## [1] 0

## Feature Selection

Finally, we decided which variables were suitable to be included in the model.

The id was removed, since it will not give us any knowledge nor be useful to predict the target.

df$customerID <- NULL

We then computed the relationship between all the variables and the target with the “catdes()” method and chose the most relevant of them for the target’s explanation.

All p-values of the Chi-squared test for categorical variables are very low, less than 0.001. The 6 variables with the lowest p-value are Contract, OnlineSecurity, TechSupport, c.tenure, InternetService, PaymentMethod. Note that the list includes a discretized numerical variable.

# Correlation between all variables and our qualitative target Churn.  
res.cat = catdes(df, grep("Churn", names(df)))  
  
# Most important categorical variables, sorted by p value  
res.cat$test.chi2

## p.value df  
## Contract 5.863038e-258 2  
## OnlineSecurity 2.661150e-185 2  
## TechSupport 1.443084e-180 2  
## c.tenure 4.192004e-178 3  
## InternetService 9.571788e-160 2  
## PaymentMethod 3.682355e-140 3  
## OnlineBackup 2.079759e-131 2  
## DeviceProtection 5.505219e-122 2  
## StreamingMovies 2.667757e-82 2  
## StreamingTV 5.528994e-82 2  
## c.MonthlyCharges 8.977393e-72 3  
## PaperlessBilling 2.614597e-58 1  
## Dependents 3.276083e-43 1  
## c.TotalCharges 3.057813e-39 3  
## SeniorCitizen 9.477904e-37 1  
## Partner 1.519037e-36 1  
## MultipleLines 3.464383e-03 2

As for numeric variables, “tenure” has the smallest p-value in the F-test, much lower than those of discrete variables. As we have already seen, there is a high correlation between “MonthlyCharges”, “tenure” and “TotalCharges” so we will only include in the models “TotalCharges” or “MonthlyCharges” together with “tenure”.

res.cat$quanti.var

## Eta2 P-value  
## tenure 0.12406504 7.999058e-205  
## TotalCharges 0.03933251 2.127212e-63  
## MonthlyCharges 0.03738671 2.706646e-60

## Profiling of the target with the selected categorical features

Lastly, we decided to make an extensive profiling of the six categorical variables that we could use in the model in order to understand them better. The main conclusions for each variable were:

* Contract: The probability of churning is decreased when the contract term increases. For example, if a costumer has a month contract and changes it to an annual the probability of not churning increases from 0.58 to 0.89.
* InternetService: People that do not have an internet service do not usually churn (7%). However, if they had a Fiber optic connection, the probability to churn increases (42%). This could be explained by the fact that users with a fast internet connection try to get the best offer for the service, but it would be necessary to make a market analysis to validate this hypothesis.
* OnlineSecurity: The probability of churning is small when the customer has online security. However, having an internet connection or not seems a more interesting feature than the variable itself, as the “No internet service” level has the smallest p-value.
* TechSupport: Having tech support increases the probability of not churning from 60% to 84% (when compared with not having it, although having internet service). Having internet service or not is, again, a more relevant feature.
* c.tenure: Loyalty is important, since people tend to churn less when they have spent longer with the service. For example, people who have spent less than 1.5 years has churned 44% of times, but only 8% of those who have stayed for more than 4.5 years have churned.
* PaymentMethod: The proportion of people that churned is very similar in all types of payment except for “Electronic check”. In this level, the proportion of churns is 45%, 18% higher than the global average.

# Global proportions of Churn categories  
proportions(table(df$Churn))

##   
## No Yes   
## 0.7346301 0.2653699

# Calculate the indexes of the variables to investigate  
names = c("Contract", "OnlineSecurity", "TechSupport", "c.tenure", "InternetService", "PaymentMethod")  
index = NULL  
  
for (i in 1:length(names)) {  
 ind = grep(names[i], colnames(df))  
 index = append(index, ind)  
}  
index = append(index, grep("Churn", names(df)))  
  
# Profiling of only those variables  
res.cat2 = catdes(df[,index], length(index))  
  
res.cat2$category

## $No  
## Cla/Mod Mod/Cla Global  
## Contract=Two year 97.16814 31.83224 24.06645  
## c.tenure=(54,72] 92.02859 32.35408 25.82706  
## InternetService=No 92.59502 27.30963 21.66690  
## TechSupport=No internet service 92.59502 27.30963 21.66690  
## OnlineSecurity=No internet service 92.59502 27.30963 21.66690  
## Contract=One year 88.73048 25.26092 20.91438  
## OnlineSecurity=Yes 85.38881 33.32045 28.66676  
## TechSupport=Yes 84.83366 33.51372 29.02172  
## PaymentMethod=Credit card (automatic) 84.75690 24.93235 21.61011  
## InternetService=DSL 81.04089 37.92037 34.37456  
## PaymentMethod=Bank transfer (automatic) 83.29016 24.85504 21.92248  
## PaymentMethod=Mailed check 80.89330 25.20294 22.88797  
## c.tenure=(36,54] 81.97970 18.72826 16.78262  
## c.tenure=(18,36] 77.29358 19.54001 18.57163  
## PaymentMethod=Electronic check 54.71459 25.00966 33.57944  
## InternetService=Fiber optic 58.10724 34.77000 43.95854  
## c.tenure=(-1,18] 55.59620 29.37766 38.81869  
## TechSupport=No 58.36453 39.17665 49.31137  
## OnlineSecurity=No 58.23328 39.36993 49.66634  
## Contract=Month-to-month 57.29032 42.90684 55.01917  
## p.value v.test  
## Contract=Two year 3.588830e-187 29.178937  
## c.tenure=(54,72] 2.745248e-113 22.620153  
## InternetService=No 6.584621e-98 20.999812  
## TechSupport=No internet service 6.584621e-98 20.999812  
## OnlineSecurity=No internet service 6.584621e-98 20.999812  
## Contract=One year 3.593041e-57 15.935502  
## OnlineSecurity=Yes 1.606459e-50 14.947938  
## TechSupport=Yes 1.323174e-46 14.334963  
## PaymentMethod=Credit card (automatic) 6.408166e-32 11.758206  
## InternetService=DSL 2.545367e-26 10.614727  
## PaymentMethod=Bank transfer (automatic) 1.180908e-24 10.250207  
## PaymentMethod=Mailed check 3.226893e-15 7.881803  
## c.tenure=(36,54] 6.217772e-14 7.503412  
## c.tenure=(18,36] 4.375264e-04 3.516348  
## PaymentMethod=Electronic check 1.790860e-136 -24.864755  
## InternetService=Fiber optic 2.289126e-148 -25.941138  
## c.tenure=(-1,18] 7.876341e-159 -26.852547  
## TechSupport=No 1.899538e-183 -28.883947  
## OnlineSecurity=No 6.171504e-190 -29.396034  
## Contract=Month-to-month 3.620915e-283 -35.959308  
##   
## $Yes  
## Cla/Mod Mod/Cla Global  
## Contract=Month-to-month 42.709677 88.550027 55.01917  
## OnlineSecurity=No 41.766724 78.170144 49.66634  
## TechSupport=No 41.635474 77.367576 49.31137  
## c.tenure=(-1,18] 44.403804 64.954521 38.81869  
## InternetService=Fiber optic 41.892765 69.395399 43.95854  
## PaymentMethod=Electronic check 45.285412 57.303371 33.57944  
## c.tenure=(18,36] 22.706422 15.890851 18.57163  
## c.tenure=(36,54] 18.020305 11.396469 16.78262  
## PaymentMethod=Mailed check 19.106700 16.479401 22.88797  
## PaymentMethod=Bank transfer (automatic) 16.709845 13.804173 21.92248  
## InternetService=DSL 18.959108 24.558587 34.37456  
## PaymentMethod=Credit card (automatic) 15.243101 12.413055 21.61011  
## TechSupport=Yes 15.166341 16.586410 29.02172  
## OnlineSecurity=Yes 14.611194 15.783842 28.66676  
## Contract=One year 11.269518 8.881755 20.91438  
## InternetService=No 7.404980 6.046014 21.66690  
## TechSupport=No internet service 7.404980 6.046014 21.66690  
## OnlineSecurity=No internet service 7.404980 6.046014 21.66690  
## c.tenure=(54,72] 7.971413 7.758159 25.82706  
## Contract=Two year 2.831858 2.568218 24.06645  
## p.value v.test  
## Contract=Month-to-month 3.620915e-283 35.959308  
## OnlineSecurity=No 6.171504e-190 29.396034  
## TechSupport=No 1.899538e-183 28.883947  
## c.tenure=(-1,18] 7.876341e-159 26.852547  
## InternetService=Fiber optic 2.289126e-148 25.941138  
## PaymentMethod=Electronic check 1.790860e-136 24.864755  
## c.tenure=(18,36] 4.375264e-04 -3.516348  
## c.tenure=(36,54] 6.217772e-14 -7.503412  
## PaymentMethod=Mailed check 3.226893e-15 -7.881803  
## PaymentMethod=Bank transfer (automatic) 1.180908e-24 -10.250207  
## InternetService=DSL 2.545367e-26 -10.614727  
## PaymentMethod=Credit card (automatic) 6.408166e-32 -11.758206  
## TechSupport=Yes 1.323174e-46 -14.334963  
## OnlineSecurity=Yes 1.606459e-50 -14.947938  
## Contract=One year 3.593041e-57 -15.935502  
## InternetService=No 6.584621e-98 -20.999812  
## TechSupport=No internet service 6.584621e-98 -20.999812  
## OnlineSecurity=No internet service 6.584621e-98 -20.999812  
## c.tenure=(54,72] 2.745248e-113 -22.620153  
## Contract=Two year 3.588830e-187 -29.178937

# Another visualization of the profiling  
#profiling(df[,index], df$Churn, "Churn")

# 6. Modeling

## Data splitting

# First, let's split the dataset into training and testing set.   
# We can consider that 70% of data will be used for training purpose.  
  
set.seed(123)  
  
sampling = sample.split(df$Churn, SplitRatio = 0.7)  
train = subset(df, sampling == TRUE)  
test = subset(df, sampling == FALSE)

## Modelling only with numerical variables.

# As we mentioned, there is a strong correlation between {tenure, MonthlyCharges}  
# and {TotalCharges}, as the second one is simply the product of first set.  
# So we will build two model with different data set, and keep the best one.  
  
m0.set1 = glm (Churn ~ tenure + MonthlyCharges, data = train, family = binomial)  
# Checking the Anova test, both variable are significant to our model.  
# Hence, we won't remove any of them.  
Anova(m0.set1, test = "LR")

## Analysis of Deviance Table (Type II tests)  
##   
## Response: Churn  
## LR Chisq Df Pr(>Chisq)   
## tenure 1071.50 1 < 2.2e-16 \*\*\*  
## MonthlyCharges 583.55 1 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

m0.set2 = glm (Churn ~ TotalCharges, data = train, family = binomial)  
  
# Checking the Bayesian criterion, the set {tenure, MonthlyCharges}   
# has much lower value. Hence, we'll choose this for further analysis.  
BIC(m0.set1, m0.set2)

## df BIC  
## m0.set1 3 4444.286  
## m0.set2 2 5504.792

# Checking possible transformation for previous model m0.set1  
  
m0.log = glm (Churn ~ tenure + log(MonthlyCharges), data = train, family = binomial)  
m0.sqrt = glm (Churn ~ sqrt(tenure) + MonthlyCharges, data = train, family = binomial)  
  
# We have tried several transformations for both variable (sqrt, log, exp, etc),  
# but BIC shows that the best model is the one with sqrt on the tenure.  
BIC (m0.set1, m0.log, m0.sqrt)

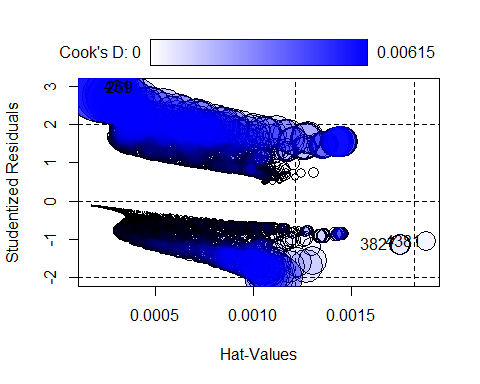
## df BIC  
## m0.set1 3 4444.286  
## m0.log 3 4465.685  
## m0.sqrt 3 4397.700

# Comparing categorical variables {c.tenure}. We'll like to make an comparison   
# between different kind of tenure (numerical and discretized)  
m1 = glm (Churn ~ c.tenure + MonthlyCharges, data = train, family = binomial)  
  
BIC(m1, m0.sqrt)

## df BIC  
## m1 5 4585.287  
## m0.sqrt 3 4397.700

# Checking the AIC and BIC parameter, we decided to keep the tenure numerical, without discretization.  
  
# We have checked alternately by changing MonthlyCharges to discretized, but  
# the AIC also get worse.

# Check the influential plot  
influent = influencePlot(m0.sqrt)[3]; influent



# Calculate D's threshold  
D\_thresh <- 2/sqrt(dim(train)[1]); D\_thresh  
  
# Cook's distance obtained from influence plot are smaller than  
# threshold, hence, we won't remove any point.

# Adding categorical variables {Contract}  
m2 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract,   
 data = train, family = binomial)  
  
# Adding {contract} indeed reduce BIC of our model.  
BIC(m0.sqrt, m2)

## df BIC  
## m0.sqrt 3 4397.700  
## m2 5 4217.893

# Adding categorical variables {InternetService}  
m3 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService,   
 data = train, family = binomial)  
  
# Adding {InternetService} indeed reduce BIC of our model.  
BIC(m2,m3)

## df BIC  
## m2 5 4217.893  
## m3 7 4190.321

We have figured out in profiling section that {InternetService} and {OnlineSecurity, TechSupport} have some level that are strongly correlated. In case that {Internet Service} = “No”, {OnlineSecurity} and {TechSupport} are also None (“No intervet service”).

Hence, to avoid dependency and NA’s features in the model, we’ll need to decide which one to keep

# Adding categorical variables {TechSupport} and {OnlineSecurity}  
m4 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + OnlineSecurity   
 + TechSupport, data = train, family = binomial)  
  
BIC(m3, m4)

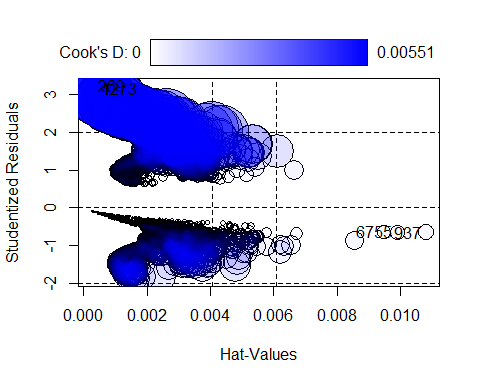
## df BIC  
## m3 7 4190.321  
## m4 8 4153.715

# The BIC criterion for m4 is smaller, but taking into account that  
# {InternetService} is more correlated with target variable {Churn}   
# and the difference of BIC is not that significant. we decided to keep  
# m3, with {InternetService}.

# Adding categorical variable {PaymentMethod}  
m5 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService  
 + PaymentMethod, data = train, family = binomial)  
  
# Adding {PaymentMethod} indeed reduce BIC of our model.  
BIC(m3, m5)

## df BIC  
## m3 7 4190.321  
## m5 10 4174.603

# Check influential plot before removing influential observation.  
influent = influencePlot(m5)[3]; influent



# Calculate D's threshold  
D\_thresh <- 2/sqrt(dim(train)[1]); D\_thresh  
  
# Cook's distance obtained from influence plot are smaller than  
# threshold, hence, we won't remove any point.

# Check all the possible interactions of model m5.  
m6 = glm (Churn ~ (sqrt(tenure) + MonthlyCharges + Contract + InternetService  
 + PaymentMethod)^2, data = train, family = binomial)  
  
# Use step function to find the combination with minimun AIC.   
step(m6)

## Start: AIC=4081.75  
## Churn ~ (sqrt(tenure) + MonthlyCharges + Contract + InternetService +   
## PaymentMethod)^2  
##   
## Df Deviance AIC  
## - Contract:PaymentMethod 6 4006.0 4076.0  
## - sqrt(tenure):InternetService 2 4000.4 4078.4  
## - sqrt(tenure):MonthlyCharges 1 3999.7 4079.7  
## - InternetService:PaymentMethod 6 4010.3 4080.3  
## - MonthlyCharges:PaymentMethod 3 4004.7 4080.7  
## <none> 3999.7 4081.7  
## - Contract:InternetService 4 4013.3 4087.3  
## - MonthlyCharges:Contract 2 4009.6 4087.6  
## - sqrt(tenure):Contract 2 4009.6 4087.6  
## - sqrt(tenure):PaymentMethod 3 4012.2 4088.2  
## - MonthlyCharges:InternetService 2 4031.4 4109.4  
##   
## Step: AIC=4076.02  
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +   
## PaymentMethod + sqrt(tenure):MonthlyCharges + sqrt(tenure):Contract +   
## sqrt(tenure):InternetService + sqrt(tenure):PaymentMethod +   
## MonthlyCharges:Contract + MonthlyCharges:InternetService +   
## MonthlyCharges:PaymentMethod + Contract:InternetService +   
## InternetService:PaymentMethod  
##   
## Df Deviance AIC  
## - sqrt(tenure):InternetService 2 4006.9 4072.9  
## - InternetService:PaymentMethod 6 4015.8 4073.8  
## - sqrt(tenure):MonthlyCharges 1 4006.0 4074.0  
## - MonthlyCharges:PaymentMethod 3 4010.9 4074.9  
## <none> 4006.0 4076.0  
## - Contract:InternetService 4 4020.0 4082.0  
## - sqrt(tenure):Contract 2 4016.5 4082.5  
## - sqrt(tenure):PaymentMethod 3 4018.6 4082.6  
## - MonthlyCharges:Contract 2 4016.6 4082.6  
## - MonthlyCharges:InternetService 2 4037.4 4103.4  
##   
## Step: AIC=4072.85  
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +   
## PaymentMethod + sqrt(tenure):MonthlyCharges + sqrt(tenure):Contract +   
## sqrt(tenure):PaymentMethod + MonthlyCharges:Contract + MonthlyCharges:InternetService +   
## MonthlyCharges:PaymentMethod + Contract:InternetService +   
## InternetService:PaymentMethod  
##   
## Df Deviance AIC  
## - sqrt(tenure):MonthlyCharges 1 4006.9 4070.9  
## - InternetService:PaymentMethod 6 4017.3 4071.3  
## - MonthlyCharges:PaymentMethod 3 4011.8 4071.8  
## <none> 4006.9 4072.9  
## - sqrt(tenure):PaymentMethod 3 4020.1 4080.1  
## - MonthlyCharges:Contract 2 4018.4 4080.4  
## - sqrt(tenure):Contract 2 4018.5 4080.5  
## - Contract:InternetService 4 4025.3 4083.3  
## - MonthlyCharges:InternetService 2 4038.1 4100.1  
##   
## Step: AIC=4070.87  
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +   
## PaymentMethod + sqrt(tenure):Contract + sqrt(tenure):PaymentMethod +   
## MonthlyCharges:Contract + MonthlyCharges:InternetService +   
## MonthlyCharges:PaymentMethod + Contract:InternetService +   
## InternetService:PaymentMethod  
##   
## Df Deviance AIC  
## - InternetService:PaymentMethod 6 4017.3 4069.3  
## - MonthlyCharges:PaymentMethod 3 4011.8 4069.8  
## <none> 4006.9 4070.9  
## - sqrt(tenure):PaymentMethod 3 4020.6 4078.6  
## - MonthlyCharges:Contract 2 4018.6 4078.6  
## - sqrt(tenure):Contract 2 4018.8 4078.8  
## - Contract:InternetService 4 4025.3 4081.3  
## - MonthlyCharges:InternetService 2 4041.3 4101.3  
##   
## Step: AIC=4069.3  
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +   
## PaymentMethod + sqrt(tenure):Contract + sqrt(tenure):PaymentMethod +   
## MonthlyCharges:Contract + MonthlyCharges:InternetService +   
## MonthlyCharges:PaymentMethod + Contract:InternetService  
##   
## Df Deviance AIC  
## - MonthlyCharges:PaymentMethod 3 4022.7 4068.7  
## <none> 4017.3 4069.3  
## - sqrt(tenure):PaymentMethod 3 4029.1 4075.1  
## - sqrt(tenure):Contract 2 4028.5 4076.5  
## - MonthlyCharges:Contract 2 4028.8 4076.8  
## - Contract:InternetService 4 4037.1 4081.1  
## - MonthlyCharges:InternetService 2 4049.3 4097.3  
##   
## Step: AIC=4068.74  
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +   
## PaymentMethod + sqrt(tenure):Contract + sqrt(tenure):PaymentMethod +   
## MonthlyCharges:Contract + MonthlyCharges:InternetService +   
## Contract:InternetService  
##   
## Df Deviance AIC  
## <none> 4022.7 4068.7  
## - sqrt(tenure):PaymentMethod 3 4031.4 4071.4  
## - sqrt(tenure):Contract 2 4033.0 4075.0  
## - MonthlyCharges:Contract 2 4035.2 4077.2  
## - Contract:InternetService 4 4042.4 4080.4  
## - MonthlyCharges:InternetService 2 4055.1 4097.1

##   
## Call: glm(formula = Churn ~ sqrt(tenure) + MonthlyCharges + Contract +   
## InternetService + PaymentMethod + sqrt(tenure):Contract +   
## sqrt(tenure):PaymentMethod + MonthlyCharges:Contract + MonthlyCharges:InternetService +   
## Contract:InternetService, family = binomial, data = train)  
##   
## Coefficients:  
## (Intercept)   
## 1.339900   
## sqrt(tenure)   
## -0.394175   
## MonthlyCharges   
## -0.018996   
## ContractOne year   
## -3.421120   
## ContractTwo year   
## -3.924823   
## InternetServiceFiber optic   
## -1.692574   
## InternetServiceNo   
## -2.818093   
## PaymentMethodCredit card (automatic)   
## -0.271243   
## PaymentMethodElectronic check   
## 0.193796   
## PaymentMethodMailed check   
## 0.090168   
## sqrt(tenure):ContractOne year   
## 0.117605   
## sqrt(tenure):ContractTwo year   
## 0.428931   
## sqrt(tenure):PaymentMethodCredit card (automatic)   
## 0.053065   
## sqrt(tenure):PaymentMethodElectronic check   
## 0.055748   
## sqrt(tenure):PaymentMethodMailed check   
## -0.104976   
## MonthlyCharges:ContractOne year   
## 0.039591   
## MonthlyCharges:ContractTwo year   
## -0.004834   
## MonthlyCharges:InternetServiceFiber optic   
## 0.041717   
## MonthlyCharges:InternetServiceNo   
## 0.064961   
## ContractOne year:InternetServiceFiber optic   
## -2.259452   
## ContractTwo year:InternetServiceFiber optic   
## -0.687971   
## ContractOne year:InternetServiceNo   
## 0.989583   
## ContractTwo year:InternetServiceNo   
## -0.411802   
##   
## Degrees of Freedom: 4929 Total (i.e. Null); 4907 Residual  
## Null Deviance: 5704   
## Residual Deviance: 4023 AIC: 4069

Df Deviance AIC

4022.7 4068.7 - sqrt(tenure):PaymentMethod 3 4031.4 4071.4 - sqrt(tenure):Contract 2 4033.0 4075.0 - MonthlyCharges:Contract 2 4035.2 4077.2 - Contract:InternetService 4 4042.4 4080.4 - MonthlyCharges:InternetService 2 4055.1 4097.1

# Check the interaction between sqrt(tenure):PaymentMethod and  
# sqrt(tenure):Contract  
m7 = glm(Churn ~ sqrt(tenure) \* PaymentMethod + sqrt(tenure) \* Contract +  
 MonthlyCharges + InternetService + PaymentMethod, data = train,   
 family = binomial)  
  
# Based on BIC criterion, no improvement is obtained.  
BIC (m5, m7)

## df BIC  
## m5 10 4174.603  
## m7 15 4194.054

# Check the interaction between MonthlyCharges and InternetService  
m8 = glm(Churn ~ sqrt(tenure) + Contract + MonthlyCharges \* InternetService   
 + PaymentMethod, data = train, family = binomial)  
  
# BIC improved from 4174 to 4164, but with the cost of 2 degree of freedom.  
BIC(m5,m8)

## df BIC  
## m5 10 4174.603  
## m8 12 4164.660

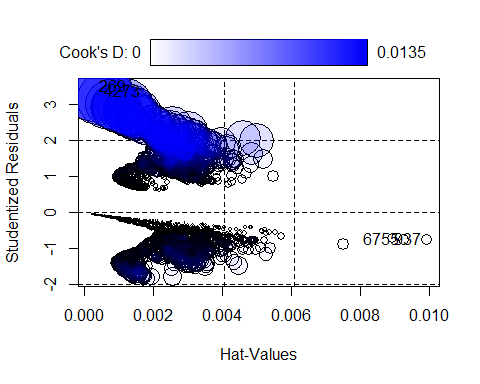
# It's a trade-off between simpleness of the model and the accuracy.  
summary(m8)

##   
## Call:  
## glm(formula = Churn ~ sqrt(tenure) + Contract + MonthlyCharges \*   
## InternetService + PaymentMethod, family = binomial, data = train)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.750286 0.297440 2.522 0.01165  
## sqrt(tenure) -0.348759 0.023305 -14.965 < 2e-16  
## ContractOne year -0.841360 0.128380 -6.554 5.61e-11  
## ContractTwo year -1.728150 0.203049 -8.511 < 2e-16  
## MonthlyCharges -0.009692 0.005254 -1.845 0.06507  
## InternetServiceFiber optic -1.504022 0.481191 -3.126 0.00177  
## InternetServiceNo -2.268409 1.568535 -1.446 0.14812  
## PaymentMethodCredit card (automatic) 0.011553 0.136266 0.085 0.93243  
## PaymentMethodElectronic check 0.453998 0.113412 4.003 6.25e-05  
## PaymentMethodMailed check -0.154366 0.138451 -1.115 0.26487  
## MonthlyCharges:InternetServiceFiber optic 0.034298 0.006722 5.103 3.35e-07  
## MonthlyCharges:InternetServiceNo 0.049096 0.075285 0.652 0.51431  
##   
## (Intercept) \*   
## sqrt(tenure) \*\*\*  
## ContractOne year \*\*\*  
## ContractTwo year \*\*\*  
## MonthlyCharges .   
## InternetServiceFiber optic \*\*   
## InternetServiceNo   
## PaymentMethodCredit card (automatic)   
## PaymentMethodElectronic check \*\*\*  
## PaymentMethodMailed check   
## MonthlyCharges:InternetServiceFiber optic \*\*\*  
## MonthlyCharges:InternetServiceNo   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5704.4 on 4929 degrees of freedom  
## Residual deviance: 4062.6 on 4918 degrees of freedom  
## AIC: 4086.6  
##   
## Number of Fisher Scoring iterations: 6

# Check the effect of linked function  
m9 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService  
 + PaymentMethod, data = train, family = binomial(link = "probit"))  
  
# Based on BIC criterion, no improvement is obtained.  
BIC(m5, m9)

## df BIC  
## m5 10 4174.603  
## m9 10 4177.774

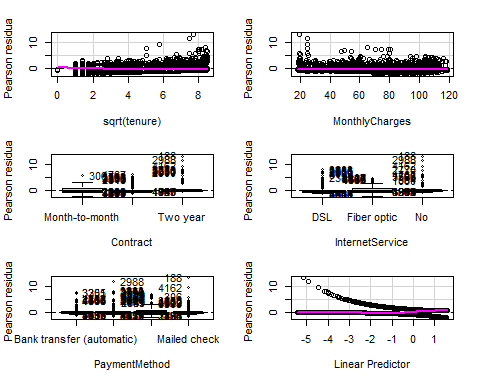
influencePlot(m9)



## StudRes Hat CookD  
## 269 3.5165320 0.0002975136 0.0134537052  
## 937 -0.7468956 0.0099201507 0.0003234289  
## 4273 3.3690041 0.0004346500 0.0119056685  
## 6755 -0.7428127 0.0092671487 0.0002981121

# Observation 269 and 4273 may be an influential point, but both of them  
# are smaller than threshold.

# Goodness of fit  
  
# Residual plot  
  
residualPlots(m5)



## Test stat Pr(>|Test stat|)   
## sqrt(tenure) 21.586 3.384e-06 \*\*\*  
## MonthlyCharges 21.877 2.907e-06 \*\*\*  
## Contract   
## InternetService   
## PaymentMethod   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Model prediction

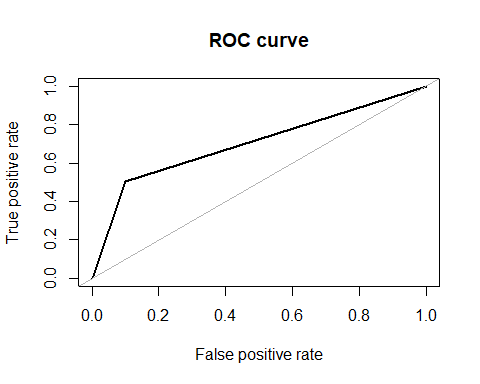
# First, we compute the probability of Churn for each observation (from test)   
# with predict function.  
predictions = predict(m5, test, type = "response")  
  
# Then, for those that have a probability higher than 0.5, we can consider  
# Churn == "Yes"  
probability = ifelse(predictions >= 0.5, "Yes", "No")  
  
# Finally, compute the Confusion Matrix of predicted result.  
CM = table(test$Churn, probability, dnn = c("Actual Churn", "Predicted Churn")); CM

## Predicted Churn  
## Actual Churn No Yes  
## No 1395 157  
## Yes 279 282

accuracy = sum(diag(CM))/dim(test)[1]\*100; accuracy

## [1] 79.36583

roc.curve(test$Churn, probability)



## Area under the curve (AUC): 0.701

library(DescTools)

##   
## Attaching package: 'DescTools'

## The following object is masked from 'package:car':  
##   
## Recode

PseudoR2(m5, which = "McFadden")

## McFadden   
## 0.2830884

# Annex