Report

Silvia Ferrer and Ignacio Lloret

2023-11-27

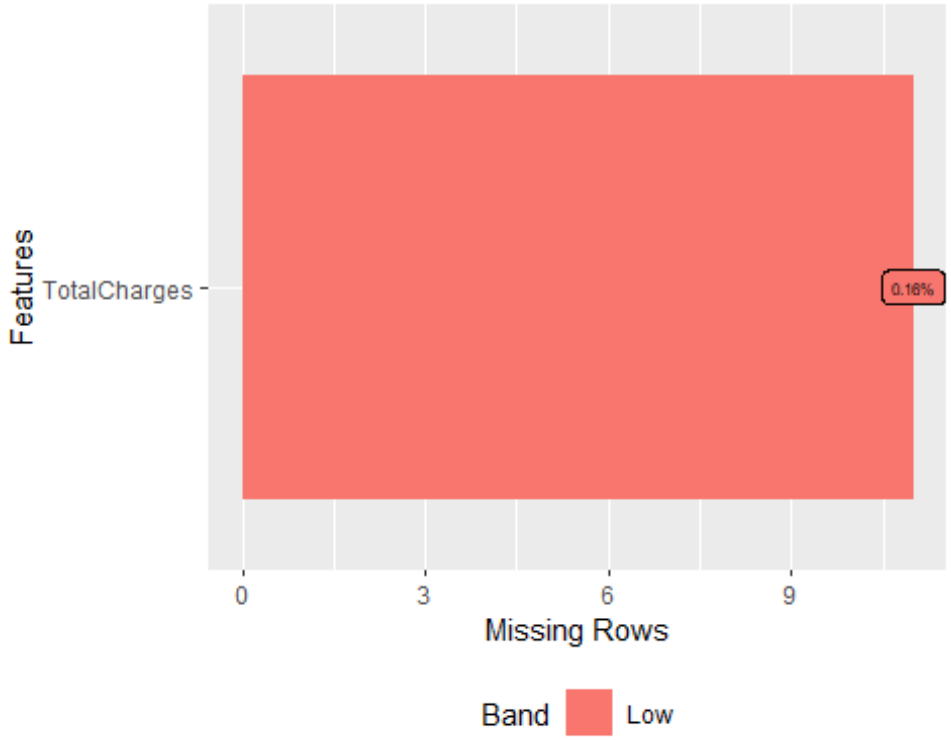
Data Preparation

Missing data and Errors

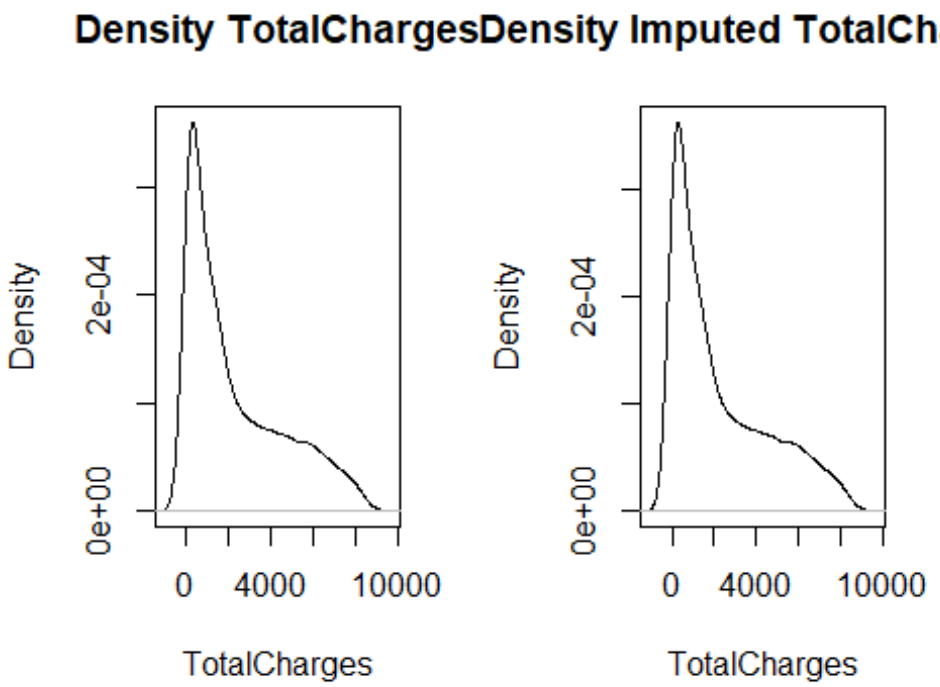
Firstly, we removed possible duplicates from the dataset using the distinct function. Then, we factorized the variable SeniorCitizen, as it only has two categories. We then excluded the customerID variable since it is a unique categorical identifier that is not useful for the model, and analyzing its data distribution does not provide meaningful insights.

Checking the missing data in the dataset and which variables have NA's we can see that the ones with missing values is TotalCharges. Investigating the observations with missing values to understand the underlying reason, we found that all these observations have tenure=0. We decided that the most appropriate option is to manually impute these TotalCharges with 0. If the tenure is 0, it implies that the contract has not started, indicating no debt or amount to be paid. We validated the imputation using density plots and confirmed that the distribution remained unchanged. Therefore, we proceeded with these imputed values.

```
#, fig.height=1.5

# Duplicates observations
df1 <- distinct(df1, .keep_all = TRUE)

# Numeric to factor SeniorCitizen
df1$SeniorCitizen <- df1$SeniorCitizen %>% as.factor()

# Take off the variable customerID
df1 <- subset(df1, select = -customerID)

cat_keep <- names(df1)[sapply(df1, function(x) is.character(x))]
numeric_columns <-  names(df1)[sapply(df1, function(x) is.numeric(x))]

df1[cat_keep] <- lapply(df1[cat_keep], as.factor) ## Create Factors
df1[numeric_columns] <- lapply(df1[numeric_columns], as.numeric)

# Missing values
plot_missing(df1, missing_only = TRUE, group = list("Low" = 0.05,
"Medium"=0.25, "High"=0.5, "Very High" =1), geom_label_args = list("size" =
2))
```




```r
observaciones_na <- df1 %>% filter(is.na(TotalCharges))
print(observaciones_na$tenure)

##  [1] 0 0 0 0 0 0 0 0 0 0 0

# Errors or inconsistencies -> imputed
df2 <- df1
df2$TotalCharges <- ifelse(is.na(df2$TotalCharges) & df2$tenure == 0, 0,
df2$TotalCharges)

# validation
summary(df2$TotalCharges)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   398.6  1394.5  2279.7  3786.6  8684.8

summary(df1$TotalCharges)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    18.8   401.4  1397.5  2283.3  3794.7  8684.8      11

par(mfrow=c(1,2))
plot(density(df1$TotalCharges,na.rm=TRUE), main = "Density TotalCharges",
     xlab = "TotalCharges", ylab = "Density")
plot(density(df2$TotalCharges,na.rm=TRUE), main = "Density Imputed
TotalCharges",
     xlab = "TotalCharges", ylab = "Density")
```




```
#inconsistencias con el No phone service o No internet service
summary(df2) # misma frec de no phone y internet service en las variables
```

```
##     gender     SeniorCitizen Partner    Dependents     tenure
PhoneService
##  Female:3488   0:5901        No :3641   No :4933   Min.   : 0.00   No :
682
##  Male  :3555   1:1142        Yes:3402   Yes:2110   1st Qu.: 9.00
Yes:6361
##                                                    Median :29.00
##                                                    Mean   :32.37
##                                                    3rd Qu.:55.00
##                                                    Max.   :72.00
##           MultipleLines     InternetService            OnlineSecurity
##  No              :3390   DSL        :2421   No                 :3498
##  No phone service: 682   Fiber optic:3096   No internet service:1526
##  Yes             :2971   No         :1526   Yes                :2019
##
##
##
##               OnlineBackup            DeviceProtection
##  No                 :3088   No                 :3095
##  No internet service:1526   No internet service:1526
##  Yes                :2429   Yes                :2422
##
##
##
```

```
##               TechSupport                StreamingTV
##  No                  :3473   No                  :2810
##  No internet service:1526   No internet service:1526
##  Yes                 :2044   Yes                 :2707
##
##
##
##             StreamingMovies           Contract    PaperlessBilling
##  No                  :2785   Month-to-month:3875   No :2872
##  No internet service:1526   One year      :1473   Yes:4171
##  Yes                 :2732   Two year      :1695
##
##
##
##                    PaymentMethod  MonthlyCharges    TotalCharges    Churn
##  Bank transfer (automatic):1544   Min.   : 18.25   Min.   :   0.0   No
:5174
##  Credit card (automatic)  :1522   1st Qu.: 35.50   1st Qu.: 398.6
Yes:1869
##  Electronic check         :2365   Median : 70.35   Median :1394.5
##  Mailed check             :1612   Mean   : 64.76   Mean   :2279.7
##                                   3rd Qu.: 89.85   3rd Qu.:3786.6
##                                   Max.   :118.75   Max.   :8684.8
```

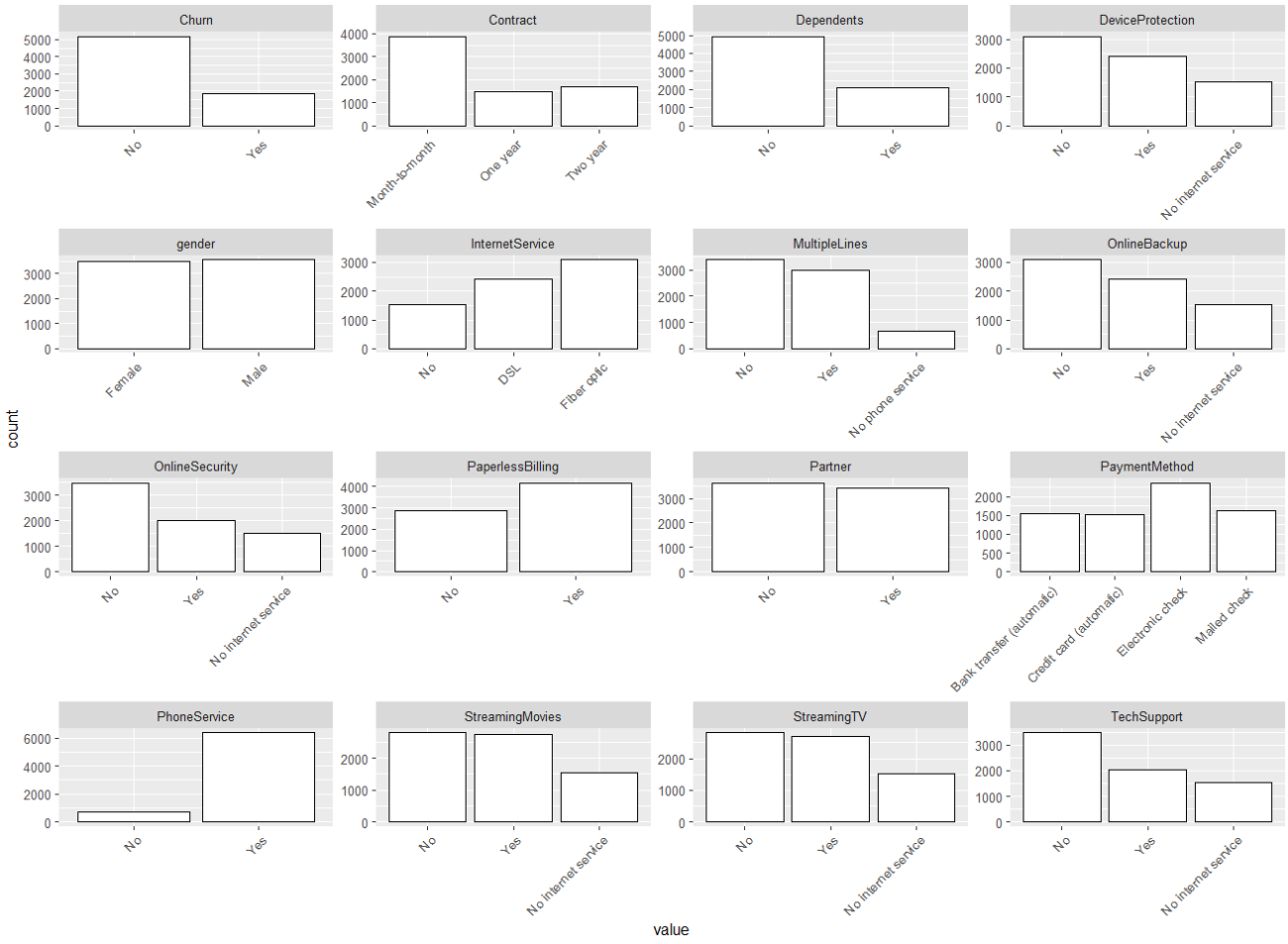
Variable analysis

Categorical values

To analyze the categorical variables, we have depicted a barplot for each of them in the figure below.

One of the most relevant observations is that our response variable, Churn, is unbalanced, with significantly more negative cases than positive ones.

```
p1 <- df2 %>%
  select(all_of(cat_keep)) %>%
  pivot_longer(cols=everything()) %>%
  ggplot(data=.) +
  geom_bar(aes(x=value), col="black", fill="white") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~name, scales="free", ncol=4)
p1
```

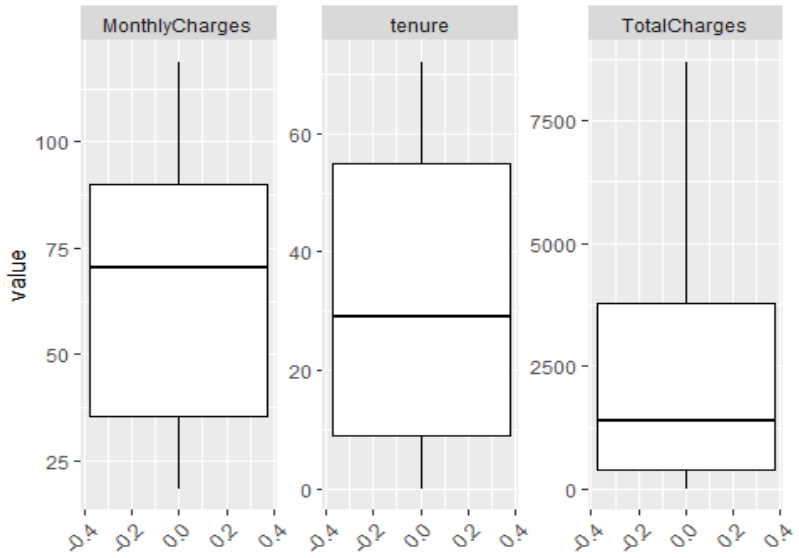


Numerical Data

In order to analyze the numerical variables, we have represented a boxplot for each of them in the figure below. Notably, none of them show univariate outliers.

Subsequently, we discretized each variable into four quartiles and represented them as factors. We displayed their frequency tables to verify that the data is appropriately distributed across each category.

```
#, fig.height=20
p2 <- df2 %>%
  select(all_of(numeric_columns)) %>%
  pivot_longer(cols=everything()) %>%
  ggplot(data=.) +
  geom_boxplot(aes(y=value), col="black", fill="white") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~name, scales="free", ncol=4)
p2
```




```
# Create a discretization of numeric variables
sm <- summary(df2$tenure)
df2$f.tenure <- ifelse(df2$tenure <= sm["1st Qu."], 1,
                ifelse(df2$tenure > sm["1st Qu."] & df2$tenure <= sm["Mean"],
2,
                ifelse(df2$tenure > sm["Mean"] & df2$tenure <= sm["3rd Qu."],
3,
                ifelse(df2$tenure > sm["3rd Qu."], 4,0))))
df2$f.tenure <- factor(df2$f.tenure,
labels=c("LowTenure","LowMidTenure","HighMidTenure","HighTenure"), order = T,
levels=c(1,2,3,4))
table(df2$f.tenure)

##
##     LowTenure  LowMidTenure HighMidTenure    HighTenure
##          1854          1921          1513          1755

sm <- summary(df2$MonthlyCharges)
df2$f.MonthlyCharges <- ifelse(df2$MonthlyCharges <= sm["1st Qu."], 1,
                ifelse(df2$MonthlyCharges > sm["1st Qu."] & df2$MonthlyCharges
<= sm["Mean"], 2,
                ifelse(df2$MonthlyCharges > sm["Mean"] & df2$MonthlyCharges <=
sm["3rd Qu."], 3,
                ifelse(df2$MonthlyCharges > sm["3rd Qu."], 4,0))))
df2$f.MonthlyCharges <- factor(df2$f.MonthlyCharges,
labels=c("LowMonthlyCharges","LowMidMonthlyCharges","HighMidMonthlyCharges","
HighMonthlyCharges"), order = T, levels=c(1,2,3,4))
table(df2$f.MonthlyCharges)

##
##     LowMonthlyCharges  LowMidMonthlyCharges HighMidMonthlyCharges
##                  1762                  1358                  2165
```

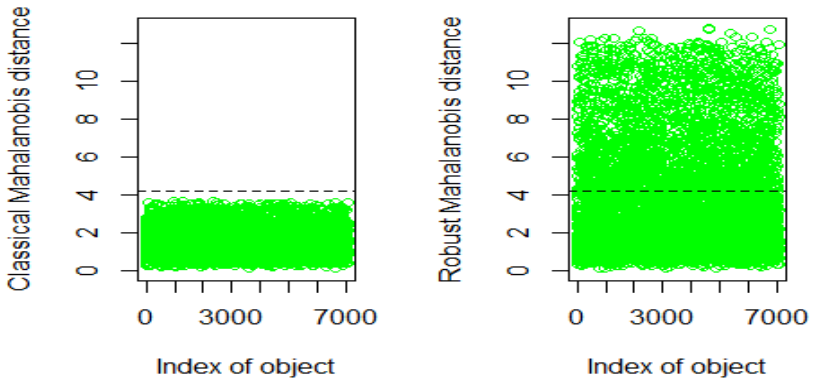
```
##    HighMonthlyCharges
##                  1758

sm <- summary(df2$TotalCharges)
df2$f.TotalCharges <- ifelse(df2$TotalCharges <= sm["1st Qu."], 1,
                ifelse(df2$TotalCharges > sm["1st Qu."] & df2$TotalCharges <=
sm["Mean"], 2,
                ifelse(df2$TotalCharges > sm["Mean"] & df2$TotalCharges <=
sm["3rd Qu."], 3,
                ifelse(df2$TotalCharges > sm["3rd Qu."], 4,0))))
df2$f.TotalCharges <- factor(df2$f.TotalCharges,
labels=c("LowTotalCharges","LowMidTotalCharges","HighMidTotalCharges","HighTo
talCharges"), order = T, levels=c(1,2,3,4))
table(df2$f.TotalCharges)

##
##     LowTotalCharges  LowMidTotalCharges HighMidTotalCharges
HighTotalCharges
##                1762                2632                 888
1761
```

Data Quality Report

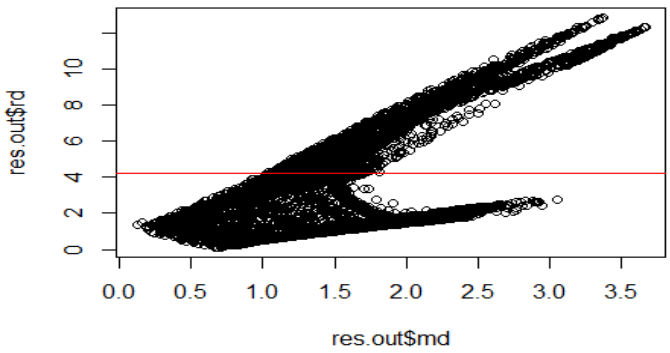
Multivariate outliers

In the initial analysis of multivariate outliers, a significance level of 0.05% was chosen as a very mild threshold. However, the vertical threshold is not visible on the graph as it extends beyond its limits. It is evident that there are no multivariate outliers beyond this threshold. We opted not to set a higher significance level because the observations are very grouped and there is no apparent clear outlier that warrants removal from the dataset.

```
df_of_interest <- df2[,c(numeric_columns)]

res.out = Moutlier(df_of_interest, quantile = 0.9995, col="green")
```



```r
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## named integer(0)

length(which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff)))

## [1] 0

par(mfrow=c(1,1))
plot( res.out$md, res.out$rd )
abline(h=res.out$cutoff, col="red")
abline(v=res.out$cutoff, col="red")
```




```r
#summary(df2[which((res.out$md > res.out$cutoff)&(res.out$rd > 
res.out$cutoff)),])
#summary(df2)

#df2 = df2[-which((res.out$md > res.out$cutoff)&(res.out$rd > 
res.out$cutoff)),]
```

Data Quality Report

As we have seen before, there are no univariate outliers, therefore, we have left the column empty, although it is represented to consider it as a parameter in the total quality sum. To measure missing values, we have conducted a column count, although we had already seen in the first section that the only column with missing values was TotalCharges, we have taken the values from the not imputed dataframe. Additionally, we consider it an error if the dataset has the tenure value equal to 0. Taking these metrics into account, we observe that the two variables with lower quality are tenure and TotalCharges. We do not believe it is necessary to look at another analysis per individuals to see the correlation with the variables because the two most related variables have been very explicitly identified in the analysis per variable.

Per variable

```r
dq <- data.frame(colnames(df1[, 1:20]))
dq$outliers <- 0
```

```r
dq$missing <- 0
dq$errors <- 0

dq$missing <- (colSums(is.na(df1[, 1:20])))
dq$errors[dq$colnames=="tenure"] <- sum(ifelse(df1$tenure == 0, 1, 0))
dq$quality <- dq$outliers + dq$missing + dq$errors
dq
```

```
##    colnames.df1...1.20.. outliers missing errors quality
## 1                 gender        0       0      0       0
## 2          SeniorCitizen        0       0      0       0
## 3                Partner        0       0      0       0
## 4             Dependents        0       0      0       0
## 5                 tenure        0       0     11      11
## 6           PhoneService        0       0      0       0
## 7          MultipleLines        0       0      0       0
## 8        InternetService        0       0      0       0
## 9         OnlineSecurity        0       0      0       0
## 10          OnlineBackup        0       0      0       0
## 11      DeviceProtection        0       0      0       0
## 12           TechSupport        0       0      0       0
## 13           StreamingTV        0       0      0       0
## 14       StreamingMovies        0       0      0       0
## 15              Contract        0       0      0       0
## 16      PaperlessBilling        0       0      0       0
## 17         PaymentMethod        0       0      0       0
## 18        MonthlyCharges        0       0      0       0
## 19          TotalCharges        0      11      0      11
## 20                 Churn        0       0      0       0
```

Profiling and Feature Selection

Interactions between the target and other variables

The results from FactoMinerR::catdes() show the relationship between the variable Churn and both categorical and quantitative variables.

For categorical variables, the chi-square test was used. The p-values for all variables are extremely small, indicating a significant association between these variables and the Churn variable. The variables with the strongest association are 'Contract', 'f.tenure', 'OnlineSecurity', and 'TechSupport', as they have the smallest p-values.

The variable Churn is also described by the categories. For the 'No' cluster, the categories with the highest v.test values (indicating a strong association) are 'Contract=Two year', 'f.tenure=HighTenure', and 'StreamingMovies=No internet service'. For the 'Yes' cluster, the categories with the highest v.test values are 'Contract=Month-to-month', 'OnlineSecurity=No', and 'TechSupport=No'.

For quantitative variables, the Eta2 statistic was used. The variable 'tenure' has the highest Eta2 value, indicating it has the strongest association with the cluster variable. The p-values for all variables are extremely small, indicating a significant association.

The variable Churn is also described by the quantitative variables. For the 'No' cluster, the variable with the highest v.test value (indicating a strong association) is 'tenure'. For the 'Yes' cluster, the variable with the highest v.test value is 'MonthlyCharges'.

As all variables are significant in relation with the variable Churn we will keep all of them at the moment.

```
catdes(df2, num.var=which(names(df2) == 'Churn'))

##
## Link between the cluster variable and the categorical variables (chi-
square test)
##
=================================================================================
====
##                          p.value df
## Contract          5.863038e-258  2
## f.tenure          1.523011e-192  3
## OnlineSecurity    2.661150e-185  2
## TechSupport       1.443084e-180  2
## InternetService   9.571788e-160  2
## PaymentMethod     3.682355e-140  3
## OnlineBackup      2.079759e-131  2
## DeviceProtection 5.505219e-122  2
## f.TotalCharges     4.965119e-85  3
## StreamingMovies    2.667757e-82  2
## StreamingTV        5.528994e-82  2
## f.MonthlyCharges   4.505436e-76  3
## PaperlessBilling   2.614597e-58  1
## Dependents         3.276083e-43  1
## SeniorCitizen      9.477904e-37  1
## Partner            1.519037e-36  1
## MultipleLines      3.464383e-03  2
##
## Description of each cluster by the categories
## ============================================
## $No
##                                              Cla/Mod  Mod/Cla   Global
## Contract=Two year                           97.16814 31.83224 24.06645
## f.tenure=HighTenure                         92.25071 31.29107 24.91836
## StreamingMovies=No internet service         92.59502 27.30963 21.66690
## StreamingTV=No internet service             92.59502 27.30963 21.66690
## TechSupport=No internet service             92.59502 27.30963 21.66690
## DeviceProtection=No internet service        92.59502 27.30963 21.66690
## OnlineBackup=No internet service            92.59502 27.30963 21.66690
## OnlineSecurity=No internet service          92.59502 27.30963 21.66690
```

```
## InternetService=No                             92.59502 27.30963 21.66690
## f.MonthlyCharges=LowMonthlyCharges              88.76277 30.22806 25.01775
## PaperlessBilling=No                             83.66992 46.44376 40.77808
## 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
## Dependents=Yes                                  84.54976 34.48009 29.95882
## f.TotalCharges=HighTotalCharges                 85.51959 29.10707 25.00355
## Partner=Yes                                     80.33510 52.82180 48.30328
## SeniorCitizen=0                                 76.39383 87.12795 83.78532
## 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
## f.tenure=HighMidTenure                          81.95638 23.96598 21.48232
## PaymentMethod=Mailed check                      80.89330 25.20294 22.88797
## OnlineBackup=Yes                                78.46851 36.83804 34.48814
## DeviceProtection=Yes                            77.49794 36.27754 34.38875
## f.TotalCharges=LowMidTotalCharges               76.02584 38.67414 37.37044
## f.MonthlyCharges=LowMidMonthlyCharges           76.73049 20.13916 19.28156
## MultipleLines=No                                74.95575 49.11094 48.13290
## MultipleLines=Yes                               71.39010 40.99343 42.18373
## StreamingMovies=Yes                             70.05857 36.99266 38.79029
## StreamingTV=Yes                                 69.92981 36.58678 38.43533
## f.MonthlyCharges=HighMonthlyCharges             67.12173 22.80634 24.96095
## StreamingTV=No                                  66.47687 36.10359 39.89777
## StreamingMovies=No                              66.31957 35.69772 39.54281
## f.MonthlyCharges=HighMidMonthlyCharges          64.11085 26.82644 30.73974
## SeniorCitizen=1                                 58.31874 12.87205 16.21468
## Partner=No                                      67.04202 47.17820 51.69672
## Dependents=No                                   68.72086 65.51991 70.04118
## PaperlessBilling=Yes                            66.43491 53.55624 59.22192
## f.TotalCharges=LowTotalCharges                  56.75369 19.32741 25.01775
## DeviceProtection=No                             60.87237 36.41283 43.94434
## OnlineBackup=No                                 60.07124 35.85234 43.84495
## PaymentMethod=Electronic check                  54.71459 25.00966 33.57944
## f.tenure=LowTenure                              50.21575 17.99382 26.32401
## InternetService=Fiber optic                     58.10724 34.77000 43.95854
## 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
## f.tenure=HighTenure                            2.648159e-111  22.417648
## StreamingMovies=No internet service             6.584621e-98  20.999812
## StreamingTV=No internet service                 6.584621e-98  20.999812
## TechSupport=No internet service                 6.584621e-98  20.999812
## DeviceProtection=No internet service            6.584621e-98  20.999812
## OnlineBackup=No internet service                6.584621e-98  20.999812
## OnlineSecurity=No internet service              6.584621e-98  20.999812
## InternetService=No                              6.584621e-98  20.999812
```

```
## f.MonthlyCharges=LowMonthlyCharges       2.427769e-71  17.859738
## PaperlessBilling=No                      1.072745e-60  16.435085
## Contract=One year                        3.593041e-57  15.935502
## OnlineSecurity=Yes                       1.606459e-50  14.947938
## TechSupport=Yes                          1.323174e-46  14.334963
## Dependents=Yes                           3.572324e-46  14.265846
## f.TotalCharges=HighTotalCharges          1.961203e-43  13.818871
## Partner=Yes                              6.170871e-37  12.696658
## SeniorCitizen=0                          3.024931e-34  12.202212
## 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
## f.tenure=HighMidTenure                   3.472392e-18   8.694866
## PaymentMethod=Mailed check               3.226893e-15   7.881803
## OnlineBackup=Yes                         3.021982e-12   6.976698
## DeviceProtection=Yes                     2.173366e-08   5.597602
## f.TotalCharges=LowMidTotalCharges        1.584501e-04   3.777438
## f.MonthlyCharges=LowMidMonthlyCharges    2.193215e-03   3.062739
## MultipleLines=No                         6.262488e-03   2.733712
## MultipleLines=Yes                        7.843169e-04  -3.358271
## StreamingMovies=Yes                      2.922571e-07  -5.128373
## StreamingTV=Yes                          1.283457e-07  -5.281193
## f.MonthlyCharges=HighMonthlyCharges      7.414051e-12  -6.849438
## StreamingTV=No                           6.049871e-27 -10.748094
## StreamingMovies=No                       1.092934e-27 -10.904833
## f.MonthlyCharges=HighMidMonthlyCharges   2.251358e-31 -11.651621
## SeniorCitizen=1                          3.024931e-34 -12.202212
## Partner=No                               6.170871e-37 -12.696658
## Dependents=No                            3.572324e-46 -14.265846
## PaperlessBilling=Yes                     1.072745e-60 -16.435085
## f.TotalCharges=LowTotalCharges           8.566779e-71 -17.789218
## DeviceProtection=No                      1.116896e-99 -21.192627
## OnlineBackup=No                         3.366400e-112 -22.509287
## PaymentMethod=Electronic check          1.790860e-136 -24.864755
## f.tenure=LowTenure                      1.176431e-143 -25.520203
## InternetService=Fiber optic             2.289126e-148 -25.941138
## 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
## InternetService=Fiber optic              41.892765 69.395399 43.95854
## f.tenure=LowTenure                       49.784250 49.384698 26.32401
## PaymentMethod=Electronic check           45.285412 57.303371 33.57944
## OnlineBackup=No                          39.928756 65.971108 43.84495
## DeviceProtection=No                      39.127625 64.794007 43.94434
```

```
## f.TotalCharges=LowTotalCharges           43.246311 40.770465 25.01775
## PaperlessBilling=Yes                     33.565092 74.906367 59.22192
## Dependents=No                            31.279140 82.557517 70.04118
## Partner=No                               32.957979 64.205457 51.69672
## SeniorCitizen=1                          41.681261 25.468165 16.21468
## f.MonthlyCharges=HighMidMonthlyCharges   35.889145 41.573034 30.73974
## StreamingMovies=No                       33.680431 50.187266 39.54281
## StreamingTV=No                           33.523132 50.401284 39.89777
## f.MonthlyCharges=HighMonthlyCharges      32.878271 30.925629 24.96095
## StreamingTV=Yes                          30.070188 43.552702 38.43533
## StreamingMovies=Yes                      29.941435 43.766720 38.79029
## MultipleLines=Yes                        28.609896 45.478866 42.18373
## MultipleLines=No                         25.044248 45.425361 48.13290
## f.MonthlyCharges=LowMidMonthlyCharges    23.269514 16.907437 19.28156
## f.TotalCharges=LowMidTotalCharges        23.974164 33.761370 37.37044
## DeviceProtection=Yes                     22.502064 29.159979 34.38875
## OnlineBackup=Yes                         21.531494 27.982879 34.48814
## PaymentMethod=Mailed check               19.106700 16.479401 22.88797
## f.tenure=HighMidTenure                   18.043622 14.606742 21.48232
## 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
## SeniorCitizen=0                          23.606168 74.531835 83.78532
## Partner=Yes                              19.664903 35.794543 48.30328
## f.TotalCharges=HighTotalCharges          14.480409 13.643660 25.00355
## Dependents=Yes                           15.450237 17.442483 29.95882
## 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
## PaperlessBilling=No                      16.330084 25.093633 40.77808
## f.MonthlyCharges=LowMonthlyCharges       11.237230 10.593900 25.01775
## StreamingMovies=No internet service       7.404980  6.046014 21.66690
## StreamingTV=No internet service           7.404980  6.046014 21.66690
## TechSupport=No internet service           7.404980  6.046014 21.66690
## DeviceProtection=No internet service      7.404980  6.046014 21.66690
## OnlineBackup=No internet service          7.404980  6.046014 21.66690
## OnlineSecurity=No internet service        7.404980  6.046014 21.66690
## InternetService=No                        7.404980  6.046014 21.66690
## f.tenure=HighTenure                       7.749288  7.276619 24.91836
## 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
## InternetService=Fiber optic              2.289126e-148  25.941138
## f.tenure=LowTenure                       1.176431e-143  25.520203
## PaymentMethod=Electronic check           1.790860e-136  24.864755
## OnlineBackup=No                          3.366400e-112  22.509287
## DeviceProtection=No                       1.116896e-99  21.192627
## f.TotalCharges=LowTotalCharges            8.566779e-71  17.789218
```

```
## PaperlessBilling=Yes                          1.072745e-60  16.435085
## Dependents=No                                 3.572324e-46  14.265846
## Partner=No                                    6.170871e-37  12.696658
## SeniorCitizen=1                               3.024931e-34  12.202212
## f.MonthlyCharges=HighMidMonthlyCharges        2.251358e-31  11.651621
## StreamingMovies=No                            1.092934e-27  10.904833
## StreamingTV=No                                6.049871e-27  10.748094
## f.MonthlyCharges=HighMonthlyCharges           7.414051e-12   6.849438
## StreamingTV=Yes                               1.283457e-07   5.281193
## StreamingMovies=Yes                           2.922571e-07   5.128373
## MultipleLines=Yes                             7.843169e-04   3.358271
## MultipleLines=No                              6.262488e-03  -2.733712
## f.MonthlyCharges=LowMidMonthlyCharges         2.193215e-03  -3.062739
## f.TotalCharges=LowMidTotalCharges             1.584501e-04  -3.777438
## DeviceProtection=Yes                          2.173366e-08  -5.597602
## OnlineBackup=Yes                              3.021982e-12  -6.976698
## PaymentMethod=Mailed check                    3.226893e-15  -7.881803
## f.tenure=HighMidTenure                        3.472392e-18  -8.694866
## 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
## SeniorCitizen=0                               3.024931e-34 -12.202212
## Partner=Yes                                   6.170871e-37 -12.696658
## f.TotalCharges=HighTotalCharges               1.961203e-43 -13.818871
## Dependents=Yes                                3.572324e-46 -14.265846
## TechSupport=Yes                               1.323174e-46 -14.334963
## OnlineSecurity=Yes                            1.606459e-50 -14.947938
## Contract=One year                             3.593041e-57 -15.935502
## PaperlessBilling=No                           1.072745e-60 -16.435085
## f.MonthlyCharges=LowMonthlyCharges            2.427769e-71 -17.859738
## StreamingMovies=No internet service           6.584621e-98 -20.999812
## StreamingTV=No internet service               6.584621e-98 -20.999812
## TechSupport=No internet service               6.584621e-98 -20.999812
## DeviceProtection=No internet service          6.584621e-98 -20.999812
## OnlineBackup=No internet service              6.584621e-98 -20.999812
## OnlineSecurity=No internet service            6.584621e-98 -20.999812
## InternetService=No                            6.584621e-98 -20.999812
## f.tenure=HighTenure                          2.648159e-111 -22.417648
## Contract=Two year                            3.588830e-187 -29.178937
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                       Eta2       P-value
## tenure          0.12406504 7.999058e-205
## TotalCharges    0.03933251  2.127212e-63
## MonthlyCharges 0.03738671  2.706646e-60
##
## Description of each cluster by quantitative variables
## ======================================================
```

```
## $No
##                    v.test Mean in category Overall mean sd in category
## tenure           29.55784         37.56997     32.37115       24.11145
## TotalCharges     16.64270       2549.91144   2279.73430     2329.72904
## MonthlyCharges -16.22582         61.26512     64.76169       31.08964
##                  Overall sd       p.value
## tenure             24.55774 5.207314e-192
## TotalCharges     2266.63354  3.418341e-62
## MonthlyCharges     30.08791  3.312724e-59
##
## $Yes
##                    v.test Mean in category Overall mean sd in category
## MonthlyCharges   16.22582         74.44133     64.76169       24.65945
## TotalCharges    -16.64270       1531.79609   2279.73430     1890.31709
## tenure          -29.55784         17.97913     32.37115       19.52590
##                  Overall sd       p.value
## MonthlyCharges     30.08791  3.312724e-59
## TotalCharges     2266.63354  3.418341e-62
## tenure             24.55774 5.207314e-192
```

Churn Modelling

Modelling using numeric variables

Initially, we built a model using only the numerical variables in our dataset. Upon examining the initial model with the vif function, we observe that there exists a high correlation between Total Charges and tenure. We will keep tenure variable, because TotalCharges is the variable that is created from tenure, in order to simplify and exclude redundant variables. Subsequent vif analysis confirmed the actual absence of multicorrelation.

Exploring interactions between these two variables gave us insignificant differences, leading us to stay with the less complex model. We tried to exchange these numeric variables with its previously created factor variables were made, but judging by the AIC parameter, the numeric variables give us better results.

Moreover, some transformations were applied to the variables. While the logarithmic transformation produced bad outcomes, the polynomial transformation significantly improved the results for tenure, although not for MonthlyCharges. Based on these findings, we kept the current best performing model which is mod_num6.

Finally, we show the effect plots of the features in the best model, and we can observe that the fewer months you stay with the company (tenure), the more likely you are to leave the company (churn yes), and the same applies in the opposite direction. Instead, the fewer monthly charges you have (MonthlyCharges), the more likely you are to stay with the company (churn no), and again, the same applies in the opposite direction.

```r
set.seed(123)
rows <- sample(nrow(df2), .75 * nrow(df2))
train_new <- df2[rows, ]
test_new <- df2[-rows, ]

## Start with the numeric variables

attach(train_new)
mod_num <- glm(Churn ~ tenure + TotalCharges + MonthlyCharges, family =
"binomial", data=train_new )
vif(mod_num) ## We can see high correlation between Total Charges and tenure.
We will keep tenure as it is the most important.

##         tenure   TotalCharges MonthlyCharges
##      13.236369      17.243623       2.293439

mod_num2 <- glm(Churn ~ tenure + MonthlyCharges, family = "binomial",
data=train_new )
vif(mod_num2) ## There is not multicorrelation

##         tenure MonthlyCharges
##       1.286659       1.286659

# Let's check if interactions may be needed

mod_num3 <- glm(Churn ~ tenure*MonthlyCharges, family="binomial",
data=train_new)
anova(mod_num2, mod_num3, test = "Chisq") # Not significant

## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges
## Model 2: Churn ~ tenure * MonthlyCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5279     4882.7
## 2      5278     4879.3  1   3.4271  0.06413 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod_num2i <- glm(Churn ~ f.tenure + f.MonthlyCharges, family = "binomial",
data=train_new )
AIC(mod_num2);AIC(mod_num2i) ## It is better with the numeric variables

## [1] 4888.737

## [1] 4981.568

mod_num4 <- glm(Churn ~ tenure + log(MonthlyCharges), family = binomial,
data=train_new)
mod_num4
```

```
##
## Call:  glm(formula = Churn ~ tenure + log(MonthlyCharges), family =
binomial,
##     data = train_new)
##
## Coefficients:
##         (Intercept)              tenure  log(MonthlyCharges)
##            -6.17416            -0.05032              1.59314
##
## Degrees of Freedom: 5281 Total (i.e. Null);  5279 Residual
## Null Deviance:       6171
## Residual Deviance: 4906  AIC: 4912
```

```
AIC(mod_num2);AIC(mod_num4) ## It is better without transformation
```

```
## [1] 4888.737
```

```
## [1] 4911.624
```

```
## Let's check for polynomial transformations
mod_num5 <- glm(Churn ~ poly(tenure,2) + poly(MonthlyCharges,2), family =
binomial, data=train_new)
summary(mod_num5)
```

```
##
## Call:
## glm(formula = Churn ~ poly(tenure, 2) + poly(MonthlyCharges,
##     2), family = binomial, data = train_new)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.37214    0.04201 -32.660  < 2e-16 ***
## poly(tenure, 2)1           -92.63576    3.63598 -25.478  < 2e-16 ***
## poly(tenure, 2)2            10.97254    2.80626   3.910 9.23e-05 ***
## poly(MonthlyCharges, 2)1  70.72197    3.26976  21.629  < 2e-16 ***
## poly(MonthlyCharges, 2)2  -0.67138    2.82535  -0.238    0.812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6171.2  on 5281  degrees of freedom
## Residual deviance: 4867.6  on 5277  degrees of freedom
## AIC: 4877.6
##
## Number of Fisher Scoring iterations: 5
```

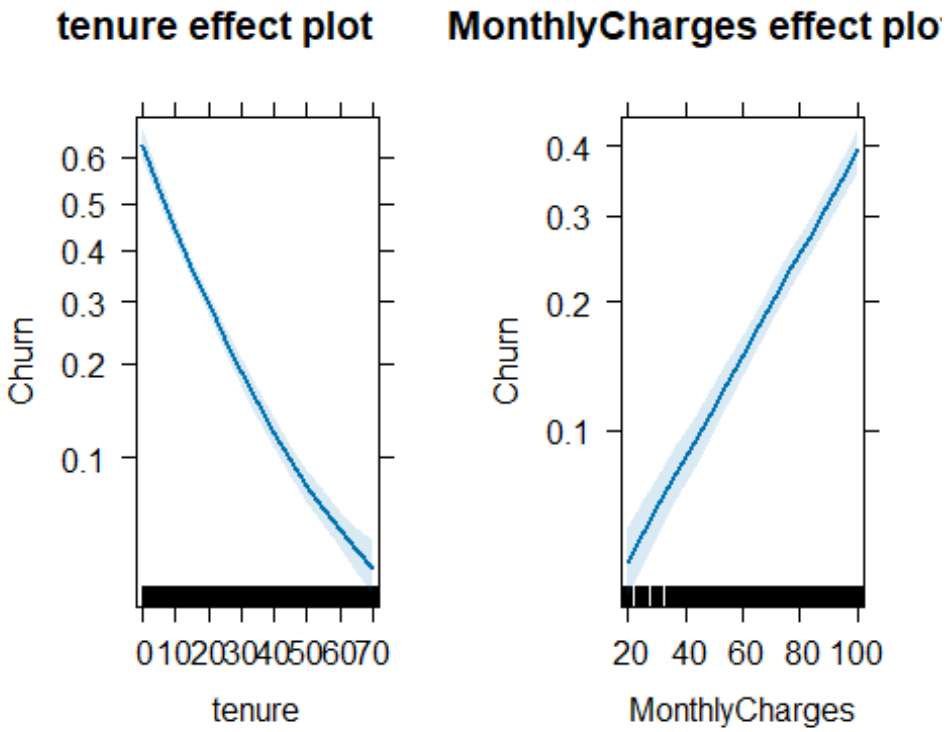
```
anova(mod_num2, mod_num5, test="Chisq") ## It is significant but
MonthlyCharges is not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges
## Model 2: Churn ~ poly(tenure, 2) + poly(MonthlyCharges, 2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5279     4882.7
## 2      5277     4867.6  2   15.139 0.0005159 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod_num6 <- glm(Churn ~ poly(tenure,2) + MonthlyCharges, family = binomial,
data=train_new)

anova(mod_num6, mod_num5, test="Chisq") ## We will keep model 6. We could try
to make polynomial of higher degrees but would be complicated to understand.
```

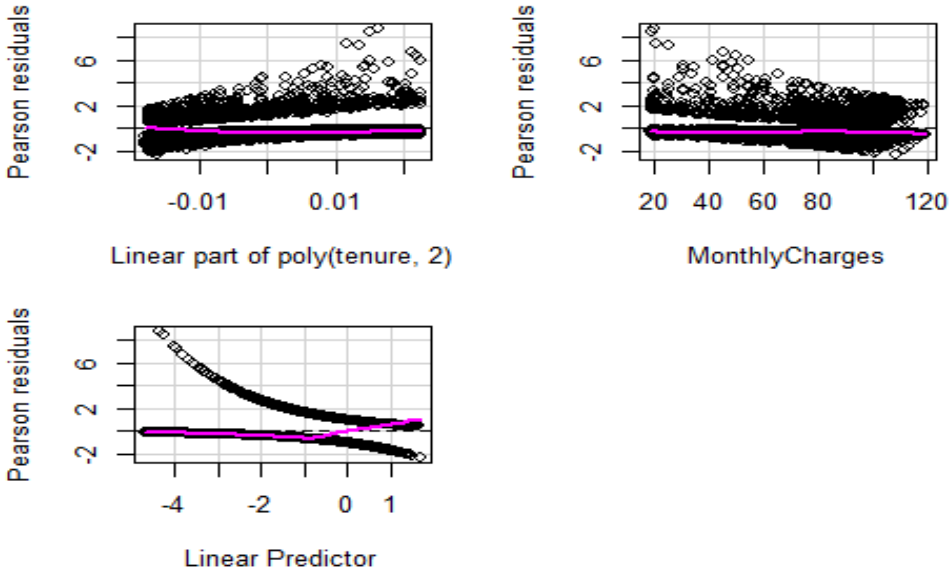
```
## Analysis of Deviance Table
##
## Model 1: Churn ~ poly(tenure, 2) + MonthlyCharges
## Model 2: Churn ~ poly(tenure, 2) + poly(MonthlyCharges, 2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5278     4867.7
## 2      5277     4867.6  1 0.056477   0.8122
```

```
plot(allEffects( mod_num6 )) ## We can see how tenure slope is smoothed in
high tenure.
```

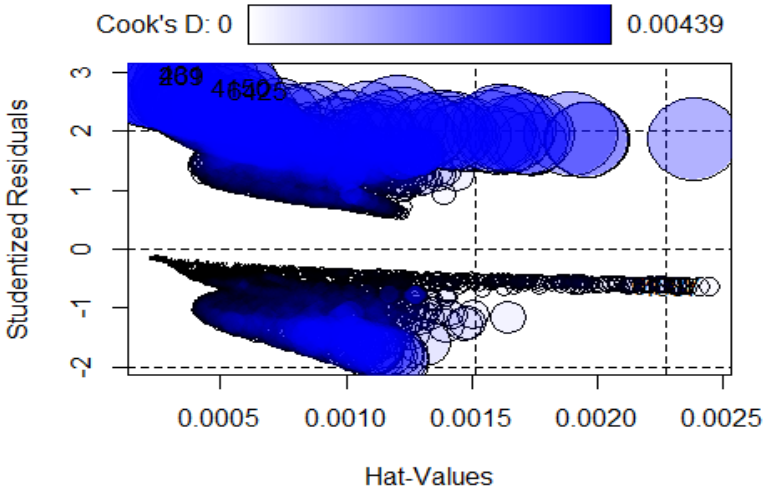


Residual analysis

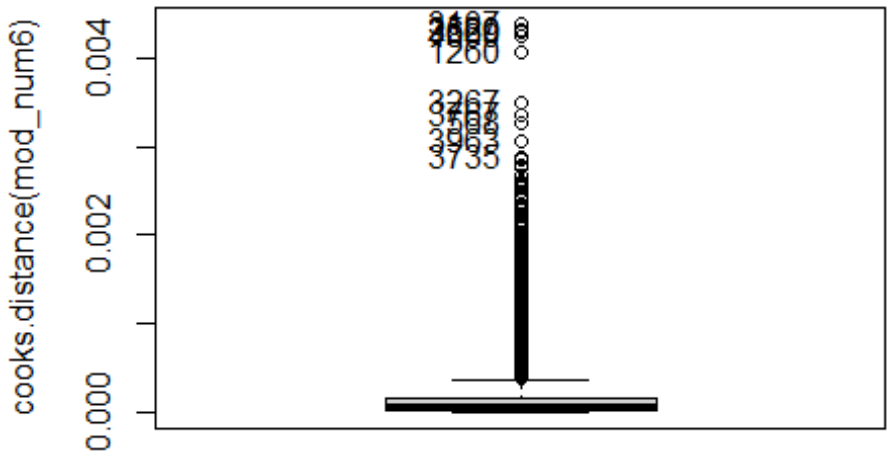
```
residualPlots( mod_num6 )
```




```
##                  Test stat Pr(>|Test stat|)
## poly(tenure, 2)
## MonthlyCharges     0.0565           0.8122

## Looks pretty flat some observations in low Monthly charges have higher
residuals but is not normal as the predictor has positive correlation, so low
Monthly charges with churn are less probable.

influencePlot( mod_num6 )
```



```
##         StudRes          Hat        CookD
## 6119 -0.6408156 0.0024069130 0.0001376051
## 269   2.9282222 0.0002283574 0.0040652746
## 4587 -0.6460006 0.0024426444 0.0001421684
## 4150  2.7261200 0.0004340039 0.0043156940
## 6425  2.6796931 0.0005017285 0.0043859131
## 431   2.9518147 0.0002246452 0.0042884924

## There are some observations that have higher residuals than expected but
are not very separate from each other. Let's check the boxplot

Boxplot(cooks.distance( mod_num6 ))
```




```
##  [1] 3107 2520 3669 4800 1260 3267 3767  568 3963 3735

## We have some influential values but it just because it is rare of low
Monthly Charges to have a churn. We believe we should keep them in the
dataset in order to not manipulate too much the model and have biased
results.
```

Adding factor main effects to the best model containing numeric variables

As a last step to create our model, we introduced all our categorical variables to the model and we run step() to remove non significant predictors. There are multiple variables that are very related with the level No Internet these generate the model to not converge in some betas. As the levels in these variables can be also categorized as No instead of No Internet Service. Also we will be able to aisle the effect of No Internet with the variable

InternetService. If more NA generate all the variance will be captured with the variable InternetService or other variable.

After refactoring all the variables that were related to each other we can see that MonthlyCharges is dependent on some of the other variables. We will remove those which are not significant and check whether we should add them or not. With the anova test we can observe that the change is not significant so we can keep the small model with the principle of parsimony. Through the vif we can also see that the multicorrelation has reduced.

Finally, we show the effect plots of the features in the model so we are able to define which category of Churn is more likely to happen when the feature takes the different values.

```
#train_new
mod <- glm(Churn ~ gender + SeniorCitizen +  Partner + Dependents +
poly(tenure, 2) + MultipleLines + InternetService + OnlineSecurity +
OnlineBackup + DeviceProtection + TechSupport + StreamingTV + StreamingMovies
+ Contract + PaperlessBilling + MonthlyCharges, data=train_new, family =
binomial)
summary(mod)

##
## Call:
## glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
##     poly(tenure, 2) + MultipleLines + InternetService + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + MonthlyCharges,
##     family = binomial, data = train_new)
##
## Coefficients: (6 not defined because of singularities)
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.18593    1.64965  -0.113 0.910259
## genderMale                           -0.05053    0.07444  -0.679 0.497271
## SeniorCitizen1                        0.23581    0.09668   2.439 0.014724
*
## PartnerYes                           -0.04925    0.08871  -0.555 0.578761
## DependentsYes                        -0.14202    0.10256  -1.385 0.166141
## poly(tenure, 2)1                    -48.46940    4.93448  -9.823  < 2e-16
***
## poly(tenure, 2)2                     23.61501    3.21735   7.340 2.14e-13
***
## MultipleLinesNo phone service        -0.07058    0.74559  -0.095 0.924582
## MultipleLinesYes                      0.46049    0.20430   2.254 0.024197
*
## InternetServiceFiber optic            1.70264    0.91938   1.852 0.064034
.
## InternetServiceNo                    -1.56283    0.92622  -1.687 0.091544
.
## OnlineSecurityNo internet service          NA         NA      NA       NA
## OnlineSecurityYes                    -0.20071    0.20616  -0.974 0.330284
```

```
## OnlineBackupNo internet service            NA         NA      NA       NA
## OnlineBackupYes                          0.01276    0.20251   0.063 0.949772
## DeviceProtectionNo internet service        NA         NA      NA       NA
## DeviceProtectionYes                      0.11152    0.20199   0.552 0.580874
## TechSupportNo internet service             NA         NA      NA       NA
## TechSupportYes                          -0.16038    0.20871  -0.768 0.442246
## StreamingTVNo internet service             NA         NA      NA       NA
## StreamingTVYes                           0.68356    0.37688   1.814 0.069720
.
## StreamingMoviesNo internet service         NA         NA      NA       NA
## StreamingMoviesYes                       0.64094    0.37613   1.704 0.088375
.
## ContractOne year                        -0.77209    0.12386  -6.233 4.56e-10
***
## ContractTwo year                        -2.01393    0.22054  -9.132  < 2e-16
***
## PaperlessBillingYes                      0.33076    0.08569   3.860 0.000113
***
## MonthlyCharges                          -0.03076    0.03651  -0.842 0.399518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6171.2  on 5281  degrees of freedom
## Residual deviance: 4429.2  on 5261  degrees of freedom
## AIC: 4471.2
##
## Number of Fisher Scoring iterations: 6

step_mod <- step(mod, trace=F)
summary(step_mod) ## There are multiple variables that are very related with
the level No Internet these generate the model to not converge in some betas.
As the levels in these variables can be also categorized as No instead of No
Internet Service. Also we will be able to aisle the effect of No Internet
with the variable InternetService. If more NA generate all the variance will
be captured with the variable InternetService or other variable.

##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + poly(tenure,
##     2) + MultipleLines + InternetService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     MonthlyCharges, family = binomial, data = train_new)
##
## Coefficients: (4 not defined because of singularities)
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -0.72477    0.58603  -1.237 0.216187
## SeniorCitizen1                      0.23123    0.09612   2.406 0.016146 *
## DependentsYes                      -0.16577    0.09360  -1.771 0.076568 .
```

```
## poly(tenure, 2)1                    -49.21233    4.80315 -10.246  < 2e-16
***
## poly(tenure, 2)2                     23.53228    3.21288   7.324 2.40e-13
***
## MultipleLinesNo phone service         0.15021    0.27613   0.544 0.586463
## MultipleLinesYes                      0.40305    0.11079   3.638 0.000275
***
## InternetServiceFiber optic            1.42884    0.31802   4.493 7.02e-06
***
## InternetServiceNo                    -1.28719    0.35466  -3.629 0.000284
***
## OnlineSecurityNo internet service          NA         NA      NA       NA
## OnlineSecurityYes                    -0.25551    0.11429  -2.236 0.025384 *
## TechSupportNo internet service             NA         NA      NA       NA
## TechSupportYes                       -0.21336    0.11783  -1.811 0.070183 .
## StreamingTVNo internet service             NA         NA      NA       NA
## StreamingTVYes                        0.57871    0.15661   3.695 0.000220
***
## StreamingMoviesNo internet service         NA         NA      NA       NA
## StreamingMoviesYes                    0.53785    0.15476   3.475 0.000510
***
## ContractOne year                     -0.76559    0.12361  -6.194 5.88e-10
***
## ContractTwo year                     -2.00258    0.22004  -9.101  < 2e-16
***
## PaperlessBillingYes                   0.32998    0.08556   3.857 0.000115
***
## MonthlyCharges                       -0.01977    0.01192  -1.659 0.097137 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6171.2  on 5281  degrees of freedom
## Residual deviance: 4430.6  on 5265  degrees of freedom
## AIC: 4464.6
##
## Number of Fisher Scoring iterations: 6
```

```
train_new$OnlineBackup    <- train_new$OnlineBackup    %>% as.character()
train_new$OnlineSecurity    <- train_new$OnlineSecurity    %>% as.character()
train_new$DeviceProtection<- train_new$DeviceProtection %>% as.character()
train_new$TechSupport     <- train_new$TechSupport     %>% as.character()
train_new$StreamingTV     <- train_new$StreamingTV     %>% as.character()
train_new$StreamingMovies <- train_new$StreamingMovies %>% as.character()

train_new$OnlineBackup     <- ifelse(train_new$OnlineBackup == 'No internet
service', 'No', train_new$OnlineBackup)
train_new$OnlineSecurity     <- ifelse(train_new$OnlineSecurity == 'No
internet service', 'No', train_new$OnlineSecurity)
```

```r
train_new$DeviceProtection <- ifelse(train_new$DeviceProtection == 'No
internet service', 'No', train_new$DeviceProtection)
train_new$TechSupport      <- ifelse(train_new$TechSupport == 'No internet
service', 'No', train_new$TechSupport)
train_new$StreamingTV      <- ifelse(train_new$StreamingTV == 'No internet
service', 'No', train_new$StreamingTV)
train_new$StreamingMovies  <- ifelse(train_new$StreamingMovies == 'No
internet service', 'No', train_new$StreamingTV)

train_new$OnlineBackup     <- train_new$OnlineBackup     %>% as.factor()
train_new$OnlineSecurity    <- train_new$OnlineSecurity    %>% as.factor()
train_new$DeviceProtection<- train_new$DeviceProtection %>% as.factor()
train_new$TechSupport      <- train_new$TechSupport      %>% as.factor()
train_new$StreamingTV      <- train_new$StreamingTV      %>% as.factor()
train_new$StreamingMovies <- train_new$StreamingMovies %>% as.factor()


mod2 <- glm(Churn ~ gender + SeniorCitizen +  Partner + Dependents +
poly(tenure, 2) + MultipleLines + InternetService + OnlineSecurity +
OnlineBackup + DeviceProtection + TechSupport + StreamingTV + Contract +
PaperlessBilling + MonthlyCharges, data=train_new, family = binomial)

summary(mod2)

##
## Call:
## glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
##     poly(tenure, 2) + MultipleLines + InternetService + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + StreamingTV +
##     Contract + PaperlessBilling + MonthlyCharges, family = binomial,
##     data = train_new)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.896568   0.440706  -6.573 4.95e-11 ***
## genderMale                      -0.052518   0.074402  -0.706 0.480274
## SeniorCitizen1                   0.235360   0.096630   2.436 0.014863 *
## PartnerYes                      -0.045129   0.088617  -0.509 0.610575
## DependentsYes                   -0.141823   0.102477  -1.384 0.166377
## poly(tenure, 2)1               -48.260650   4.930798  -9.788  < 2e-16 ***
## poly(tenure, 2)2                23.628029   3.216172   7.347 2.03e-13 ***
## MultipleLinesNo phone service    1.141370   0.225463   5.062 4.14e-07 ***
## MultipleLinesYes                 0.160485   0.103575   1.549 0.121271
## InternetServiceFiber optic       0.197162   0.253541   0.778 0.436784
## InternetServiceNo               -0.057530   0.279600  -0.206 0.836981
## OnlineSecurityYes               -0.503244   0.105288  -4.780 1.76e-06 ***
## OnlineBackupYes                 -0.289194   0.098300  -2.942 0.003262 **
## DeviceProtectionYes             -0.183167   0.104585  -1.751 0.079883 .
## TechSupportYes                  -0.463910   0.109343  -4.243 2.21e-05 ***
## StreamingTVYes                   0.090824   0.144934   0.627 0.530884
```

```
## ContractOne year               -0.771337   0.123828  -6.229 4.69e-10 ***
## ContractTwo year               -2.009535   0.220525  -9.113  < 2e-16 ***
## PaperlessBillingYes             0.336985   0.085581   3.938 8.23e-05 ***
## MonthlyCharges                  0.029601   0.008872   3.336 0.000848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6171.2  on 5281  degrees of freedom
## Residual deviance: 4432.1  on 5262  degrees of freedom
## AIC: 4472.1
##
## Number of Fisher Scoring iterations: 6
```

```
vif(mod2)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## gender            1.002976  1        1.001487
## SeniorCitizen     1.145921  1        1.070477
## Partner           1.367963  1        1.169600
## Dependents        1.274307  1        1.128852
## poly(tenure, 2)   2.674350  2        1.278806
## MultipleLines     5.631266  2        1.540464
## InternetService  26.960139  2        2.278665
## OnlineSecurity    1.337421  1        1.156469
## OnlineBackup      1.530548  1        1.237153
## DeviceProtection  1.710532  1        1.307873
## TechSupport       1.464988  1        1.210367
## StreamingTV       3.677450  1        1.917668
## Contract          1.779049  2        1.154907
## PaperlessBilling  1.127854  1        1.062005
## MonthlyCharges   42.016372  1        6.482004
```

After refactoring all the variables that were related to each other we can see that MonthlyCharges is dependent on some of the other variables. We will remove those which are not significant and check whether we should add them or not.

```
Anova(mod2, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                   LR Chisq Df Pr(>Chisq)
## gender              0.498  1  0.4802756
## SeniorCitizen       5.920  1  0.0149664 *
## Partner             0.259  1  0.6106298
## Dependents          1.922  1  0.1656116
## poly(tenure, 2)   226.114  2  < 2.2e-16 ***
```

```
## MultipleLines      33.592  2  5.076e-08 ***
## InternetService     0.790  2  0.6736513
## OnlineSecurity     23.205  1  1.456e-06 ***
## OnlineBackup        8.671  1  0.0032335 **
## DeviceProtection    3.070  1  0.0797302 .
## TechSupport        18.222  1  1.966e-05 ***
## StreamingTV         0.393  1  0.5308566
## Contract          114.051  2  < 2.2e-16 ***
## PaperlessBilling   15.596  1  7.840e-05 ***
## MonthlyCharges     11.194  1  0.0008207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod3 <- glm(Churn ~ SeniorCitizen +  poly(tenure, 2) + MultipleLines +
OnlineSecurity + OnlineBackup + TechSupport + Contract + PaperlessBilling +
MonthlyCharges, data=train_new, family = binomial)
Anova(mod3, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                   LR Chisq Df Pr(>Chisq)
## SeniorCitizen        7.795  1  0.0052386 **
## poly(tenure, 2)    263.346  2  < 2.2e-16 ***
## MultipleLines       64.374  2  1.050e-14 ***
## OnlineSecurity      33.037  1  9.043e-09 ***
## OnlineBackup        12.465  1  0.0004146 ***
## TechSupport         27.842  1  1.316e-07 ***
## Contract           130.953  2  < 2.2e-16 ***
## PaperlessBilling    17.011  1  3.716e-05 ***
## MonthlyCharges     298.361  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

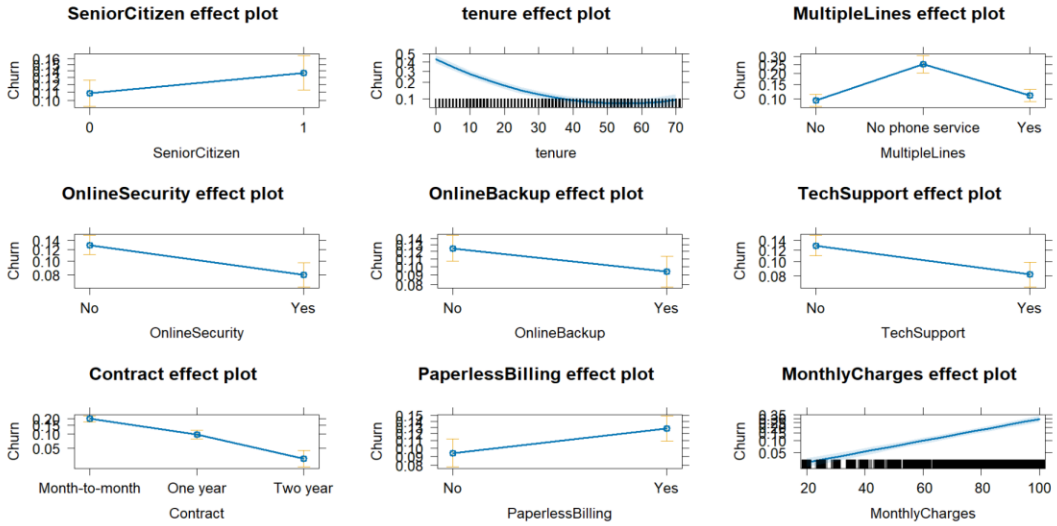
```
anova(mod3, mod2, test="Chisq") ## It is not significant so we can keep the
small model with the principle of parsimony.
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity +
##     OnlineBackup + TechSupport + Contract + PaperlessBilling +
##     MonthlyCharges
## Model 2: Churn ~ gender + SeniorCitizen + Partner + Dependents +
poly(tenure,
##     2) + MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
##     DeviceProtection + TechSupport + StreamingTV + Contract +
##     PaperlessBilling + MonthlyCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5269     4442.2
## 2      5262     4432.1  7   10.097   0.1832
```

```
vif(mod3) ## The multicorrelation has reduced.
```

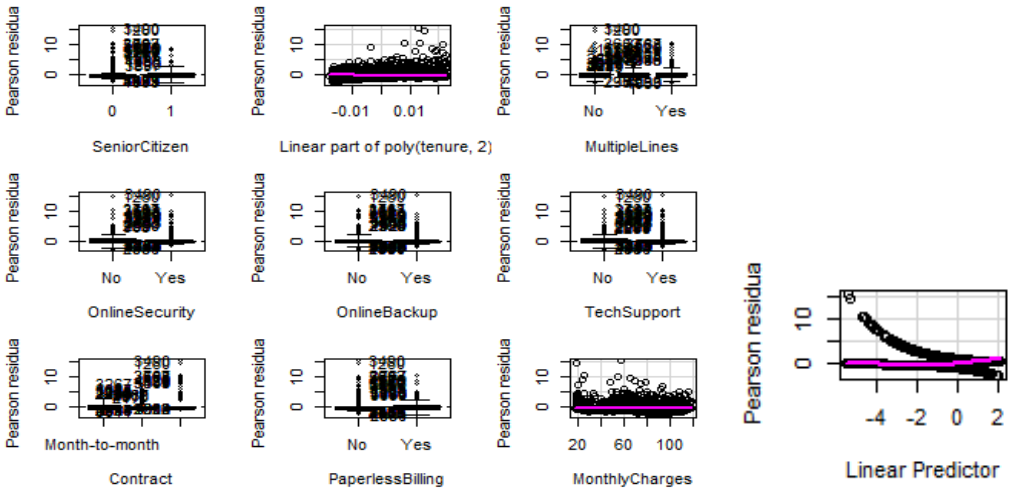
```
##                       GVIF Df GVIF^(1/(2*Df))
## SeniorCitizen     1.096933  1        1.047346
## poly(tenure, 2)   2.432127  2        1.248811
## MultipleLines     1.922042  2        1.177445
## OnlineSecurity    1.098729  1        1.048203
## OnlineBackup      1.260452  1        1.122698
## TechSupport       1.164474  1        1.079108
## Contract          1.698638  2        1.141629
## PaperlessBilling  1.121320  1        1.058924
## MonthlyCharges    2.260157  1        1.503382
```

```
plot(allEffects(mod3))
```



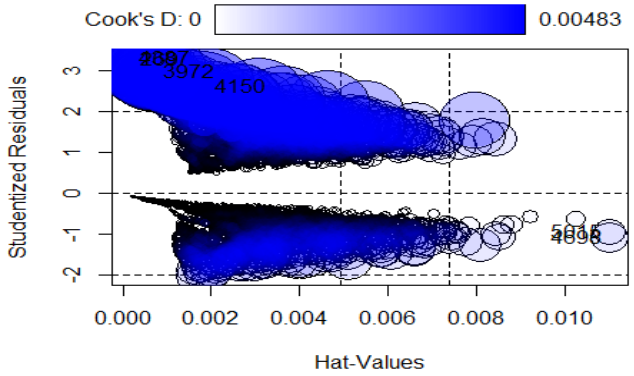
Residual analysis with categorical variables

```
residualPlots(mod3)
```



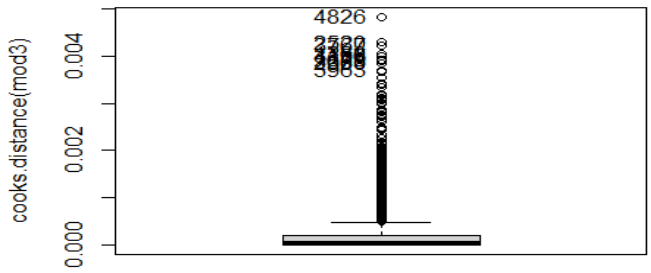
```
##                   Test stat Pr(>|Test stat|)
## SeniorCitizen
## poly(tenure, 2)
## MultipleLines
## OnlineSecurity
## OnlineBackup
## TechSupport
## Contract
## PaperlessBilling
## MonthlyCharges      0.4256           0.5142
```

```
influencePlot(mod3)
```




```
##        StudRes          Hat        CookD
## 5015 -0.954685 0.0109996750 0.0004950570
## 269   3.290572 0.0002419871 0.0040538545
## 4150  2.614033 0.0019388313 0.0042864608
## 4387  3.318092 0.0002008891 0.0036955404
## 4698 -1.093540 0.0110005478 0.0007008811
## 3972  2.989649 0.0007499970 0.0048288428
```

```
cook <- Boxplot(cooks.distance(mod3))
```




```
cookd <- sort(cooks.distance(mod3)[cook], decreasing=TRUE)
cookd
```

```
##        3972        4150        4273         269        6425        6814
## 0.004828843 0.004286461 0.004231198 0.004053854 0.004003822 0.003998106
##        6725        5590        4528        4514
## 0.003920864 0.003884925 0.003849025 0.003696433
```

```
length(rownames(train_new) %in% names(cookd)) #[1] 5282
```

```
## [1] 5282
```

Factor interactions

Now, we are searching for interactions between factors in the model, beginning by testing some combinations of variables that had sense for us to have relation between them. We identify the one that yields the best results. But, given the high quantity of variables, manually exploring combinations becomes impractical. Hence, we employ the iterative stepwise method to check different combinations. The iteration providing the best results includes interactions between OnlineSecurity and TechSupport with a high representation, and MultipleLines and TechSupport with minimal representation. We tested the one with more representation alone, and then with both interactions to assess any significant improvement. However, there is no significant change observed, leading us to choose the simpler model, mod7.

```
mod4 <- glm(Churn ~ SeniorCitizen +  poly(tenure, 2) + MultipleLines +
OnlineSecurity + OnlineBackup + TechSupport + Contract * PaperlessBilling +
MonthlyCharges, data=train_new, family = binomial)
```

```
anova(mod3, mod4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity +
##     OnlineBackup + TechSupport + Contract + PaperlessBilling +
##     MonthlyCharges
## Model 2: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity +
##     OnlineBackup + TechSupport + Contract * PaperlessBilling +
##     MonthlyCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5269     4442.2
## 2      5267     4441.7  2  0.54682   0.7608
```

```
mod5 <- glm(Churn ~ SeniorCitizen +  poly(tenure, 2) + (MultipleLines +
OnlineSecurity + OnlineBackup + TechSupport)*MonthlyCharges + Contract +
PaperlessBilling, data=train_new, family = binomial)
```

```
anova(mod3, mod5, test="Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity +
##     OnlineBackup + TechSupport + Contract + PaperlessBilling +
##     MonthlyCharges
## Model 2: Churn ~ SeniorCitizen + poly(tenure, 2) + (MultipleLines +
OnlineSecurity +
##     OnlineBackup + TechSupport) * MonthlyCharges + Contract +
##     PaperlessBilling
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5269     4442.2
## 2      5264     4437.4  5   4.7964   0.4412
```

```
mod6 <- glm(Churn ~ SeniorCitizen +  poly(tenure, 2) + (MultipleLines +
OnlineSecurity + OnlineBackup + TechSupport + MonthlyCharges)^2 + Contract +
PaperlessBilling, data=train_new, family = binomial)

step_mod <- step(mod6, trace=F) # muchas variables como para ver sus
combinaciones por eso utilizamos el metodo iterativo stepwise
summary(step_mod)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
##     OnlineSecurity + OnlineBackup + TechSupport + MonthlyCharges +
##     Contract + PaperlessBilling + MultipleLines:TechSupport +
##     OnlineSecurity:TechSupport, family = binomial, data = train_new)
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                   -3.145537   0.157115 -20.021
## SeniorCitizen1                                 0.259451   0.094679   2.740
## poly(tenure, 2)1                             -50.582871   4.767564 -10.610
## poly(tenure, 2)2                              23.668432   3.226857   7.335
## MultipleLinesNo phone service                  1.289467   0.158374   8.142
## MultipleLinesYes                               0.130584   0.104482   1.250
## OnlineSecurityYes                             -0.681131   0.113830  -5.984
## OnlineBackupYes                               -0.322770   0.089409  -3.610
## TechSupportYes                                -0.631772   0.156778  -4.030
## MonthlyCharges                                 0.033872   0.002073  16.341
## ContractOne year                              -0.817754   0.121860  -6.711
## ContractTwo year                              -2.145683   0.222457  -9.645
## PaperlessBillingYes                            0.345041   0.085413   4.040
## MultipleLinesNo phone service:TechSupportYes  -0.527034   0.320468  -1.645
## MultipleLinesYes:TechSupportYes                0.092769   0.193980   0.478
## OnlineSecurityYes:TechSupportYes               0.477683   0.202415   2.360
##                                               Pr(>|z|)
## (Intercept)                                    < 2e-16 ***
## SeniorCitizen1                                0.006138 **
## poly(tenure, 2)1                               < 2e-16 ***
## poly(tenure, 2)2                              2.22e-13 ***
```

```
## MultipleLinesNo phone service                 3.89e-16 ***
## MultipleLinesYes                              0.211365
## OnlineSecurityYes                             2.18e-09 ***
## OnlineBackupYes                               0.000306 ***
## TechSupportYes                                5.58e-05 ***
## MonthlyCharges                                 < 2e-16 ***
## ContractOne year                              1.94e-11 ***
## ContractTwo year                               < 2e-16 ***
## PaperlessBillingYes                           5.35e-05 ***
## MultipleLinesNo phone service:TechSupportYes 0.100058
## MultipleLinesYes:TechSupportYes               0.632478
## OnlineSecurityYes:TechSupportYes              0.018279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6171.2  on 5281  degrees of freedom
## Residual deviance: 4432.6  on 5266  degrees of freedom
## AIC: 4464.6
##
## Number of Fisher Scoring iterations: 6
```

```
mod7 <- glm(Churn ~ SeniorCitizen +  poly(tenure, 2) + MultipleLines +
OnlineSecurity*TechSupport + OnlineBackup + MonthlyCharges + Contract +
PaperlessBilling, data=train_new, family = binomial)

anova(mod3, mod7, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity +
##     OnlineBackup + TechSupport + Contract + PaperlessBilling +
##     MonthlyCharges
## Model 2: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity *
##     TechSupport + OnlineBackup + MonthlyCharges + Contract +
##     PaperlessBilling
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5269     4442.2
## 2      5268     4436.7  1   5.5334  0.01866 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

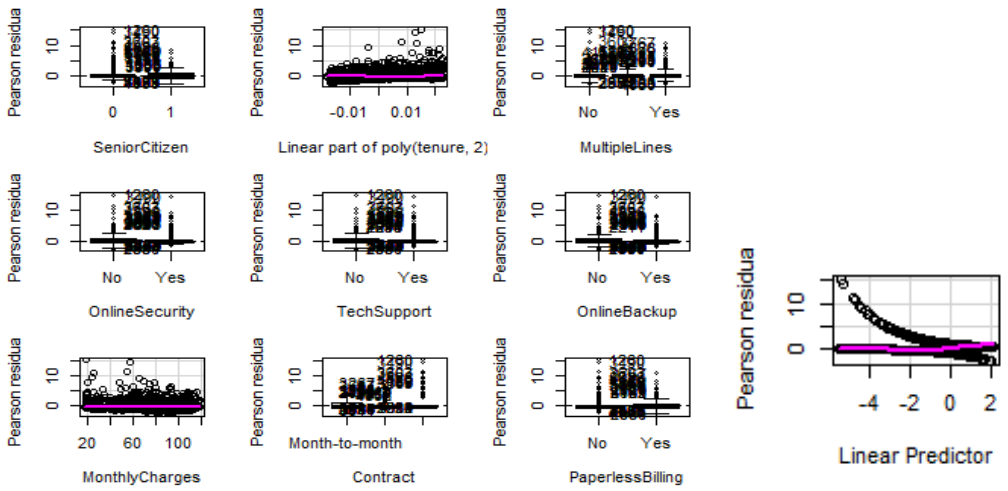
```
mod8 <- glm(Churn ~ SeniorCitizen +  poly(tenure, 2) + (MultipleLines +
OnlineSecurity)*TechSupport + OnlineBackup + MonthlyCharges + Contract +
PaperlessBilling, data=train_new, family = binomial)
```

```
anova(mod7, mod8, test="Chisq") # no cambio significant, nos quedamos con el simple mod7
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
OnlineSecurity *
##     TechSupport + OnlineBackup + MonthlyCharges + Contract +
##     PaperlessBilling
## Model 2: Churn ~ SeniorCitizen + poly(tenure, 2) + (MultipleLines +
OnlineSecurity) *
##     TechSupport + OnlineBackup + MonthlyCharges + Contract +
##     PaperlessBilling
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5268     4436.7
## 2      5266     4432.6  2    4.068   0.1308
```

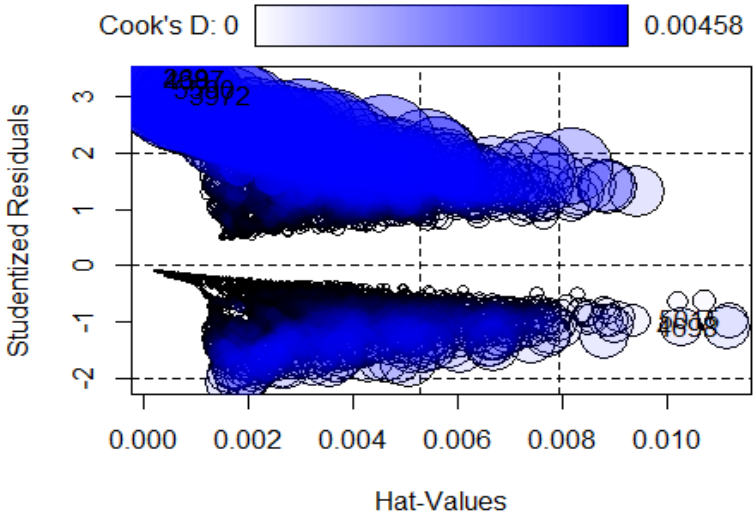
Residual analysis with the best model

```
residualPlots(mod7)
```



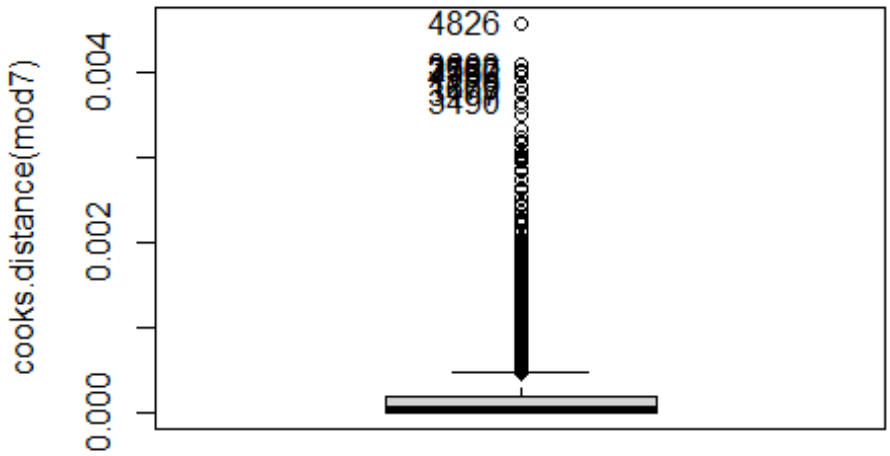

```
##                   Test stat Pr(>|Test stat|)
## SeniorCitizen
## poly(tenure, 2)
## MultipleLines
## OnlineSecurity
## TechSupport
## OnlineBackup
## MonthlyCharges      0.2991           0.5844
## Contract
## PaperlessBilling
```

```
influencePlot(mod7)
```




```
##          StudRes          Hat        CookD
## 5015 -0.9669999 0.0111716287 0.0004821169
## 269   3.3050018 0.0002356469 0.0038429673
## 4387  3.2718677 0.0002489926 0.0036447601
## 5590  3.1264571 0.0004481656 0.0040966086
## 4698 -1.1103032 0.0111496498 0.0006869491
## 3972  3.0060673 0.0007301710 0.0045848121
```

```
cook <- Boxplot(cooks.distance(mod7))
```




```
cookd <- sort(cooks.distance(mod7)[cook], decreasing=TRUE)
cookd
```

```
##        3972        5590        4528        4150        4273        6814
## 0.004584812 0.004096609 0.004038624 0.004026787 0.004008097 0.003957548
```

```
##         269        6725        6425        4387
## 0.003842967 0.003832726 0.003761736 0.003644760
```

```
length(rownames(train_new) %in% names(cookd)) #[1] 5282
```

```
## [1] 5282
```

Model Interpretation and residual analysis

```
summary(mod7)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + poly(tenure, 2) + MultipleLines +
##     OnlineSecurity * TechSupport + OnlineBackup + MonthlyCharges +
##     Contract + PaperlessBilling, family = binomial, data = train_new)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -3.142762   0.156737 -20.051  < 2e-16
***
## SeniorCitizen1                   0.261661   0.094745   2.762 0.005749 **
## poly(tenure, 2)1               -50.095970   4.742029 -10.564  < 2e-16
***
## poly(tenure, 2)2                23.855515   3.211963   7.427 1.11e-13
***
## MultipleLinesNo phone service    1.163172   0.141899   8.197 2.46e-16
***
## MultipleLinesYes                 0.149771   0.094602   1.583 0.113384
## OnlineSecurityYes               -0.682007   0.113778  -5.994 2.04e-09
***
## TechSupportYes                  -0.646185   0.114362  -5.650 1.60e-08
***
## OnlineBackupYes                 -0.321563   0.089315  -3.600 0.000318
***
## MonthlyCharges                   0.033880   0.002066  16.395  < 2e-16
***
## ContractOne year                -0.818931   0.121691  -6.730 1.70e-11
***
## ContractTwo year                -2.152023   0.221340  -9.723  < 2e-16
***
## PaperlessBillingYes              0.350968   0.085270   4.116 3.86e-05
***
## OnlineSecurityYes:TechSupportYes   0.477552   0.202089   2.363 0.018124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6171.2  on 5281  degrees of freedom
## Residual deviance: 4436.7  on 5268  degrees of freedom
```

```
## AIC: 4464.7
##
## Number of Fisher Scoring iterations: 6
```

```
vif(mod7)
```

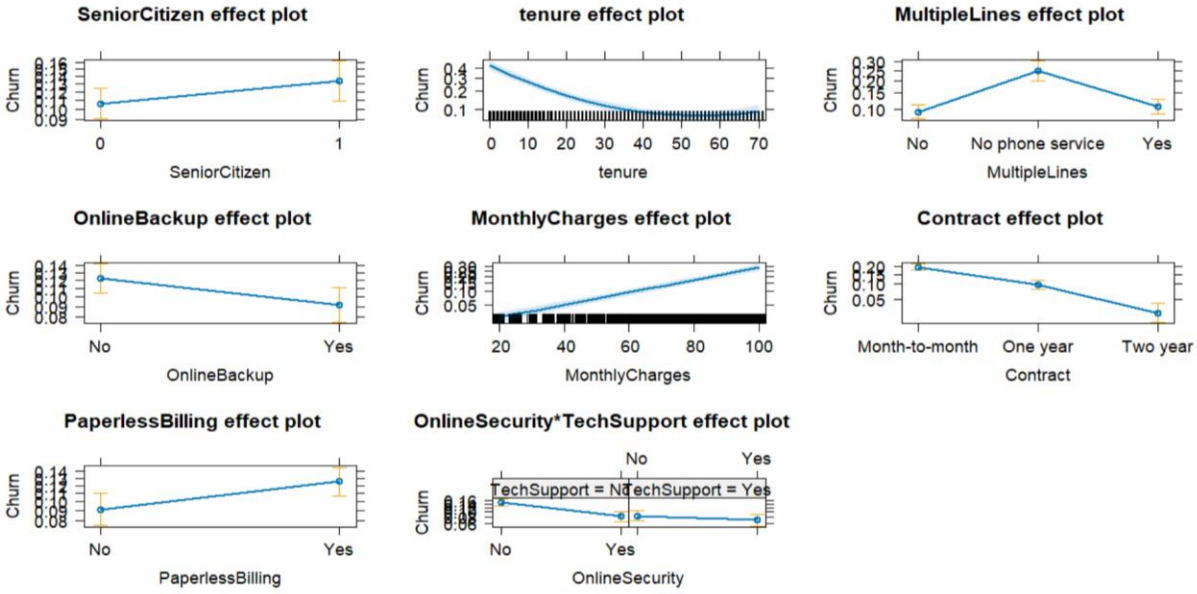
```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## SeniorCitizen              1.097232  1        1.047489
## poly(tenure, 2)            2.448750  2        1.250939
## MultipleLines              1.924345  2        1.177798
## OnlineSecurity             1.586071  1        1.259393
## TechSupport                1.625500  1        1.274951
## OnlineBackup               1.264149  1        1.124344
## MonthlyCharges             2.270735  1        1.506896
## Contract                   1.754205  2        1.150854
## PaperlessBilling           1.122019  1        1.059254
## OnlineSecurity:TechSupport 2.158190  1        1.469078
```

```
exp(mod7$coefficients)
```

```
##                       (Intercept)                    SeniorCitizen1
##                      4.316340e-02                      1.299086e+00
##                  poly(tenure, 2)1                 poly(tenure, 2)2
##                      1.752252e-22                      2.292549e+10
##    MultipleLinesNo phone service                 MultipleLinesYes
##                      3.200069e+00                      1.161568e+00
##                 OnlineSecurityYes                   TechSupportYes
##                      5.056011e-01                      5.240411e-01
##                   OnlineBackupYes                   MonthlyCharges
##                      7.250153e-01                      1.034460e+00
##                  ContractOne year                 ContractTwo year
##                      4.409028e-01                      1.162488e-01
##              PaperlessBillingYes OnlineSecurityYes:TechSupportYes
##                      1.420442e+00                      1.612123e+00
```

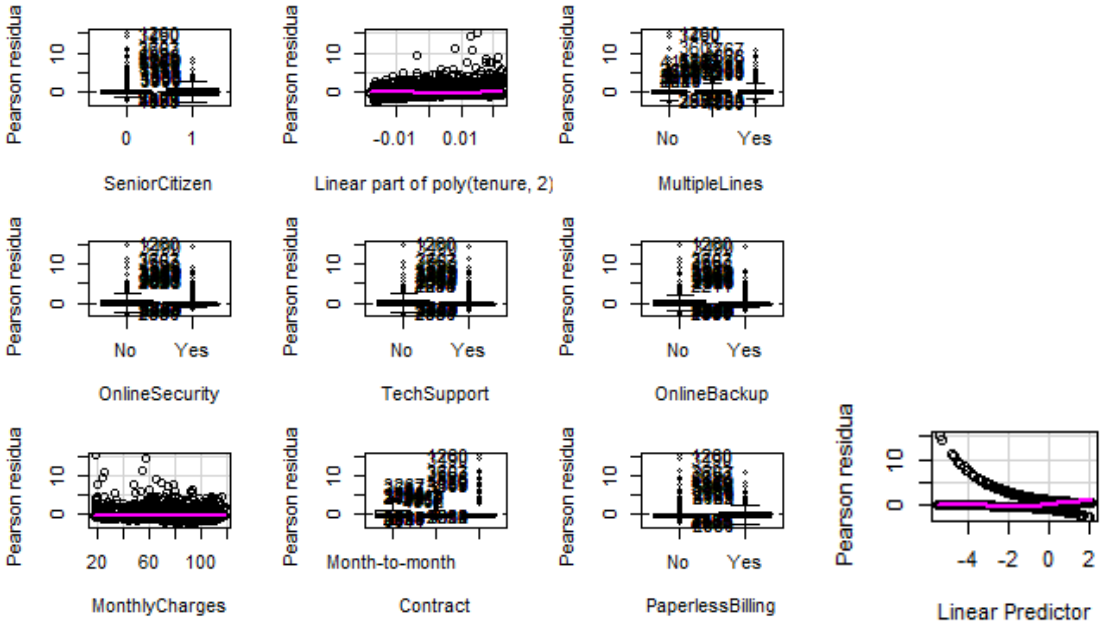
```
par(mfrow=c(1,2))
plot(allEffects(mod = mod7))
```




```
#avPlots(mod7) #es como el effect
sum( resid( mod7, "pearson") ^2 )
```

```
## [1] 5607.608
```

```
residualPlots(mod7)
```




```
##                   Test stat Pr(>|Test stat|)
## SeniorCitizen
## poly(tenure, 2)
## MultipleLines
## OnlineSecurity
```

```
## TechSupport
## OnlineBackup
## MonthlyCharges      0.2991           0.5844
## Contract
## PaperlessBilling
```

The final model has this form.

$$
\begin{aligned}
y = \quad & -3.1427\alpha + \beta_{\text{SeniorCitizen1}}0.26 + \beta_{\text{tenure}}(-50.09) + \beta_{\text{tenure}^2}23.85 \\
& + \beta_{\text{MultipleLinesNo phone service}}1.16 + \beta_{\text{MultipleLinesYes}}0.14 + \beta_{\text{OnlineSecurityYes}}(-0.68) \\
& + \beta_{\text{TechSupportYes}}(-0.64) + \beta_{\text{OnlineBackupYes}}(-0.32) + \beta_{\text{MonthlyCharges}}0.03 \\
& + \beta_{\text{ContractOne year}}(-0.81) + \beta_{\text{ContractTwo year}}(-2.15) + \beta_{\text{PaperlessBillingYes}}0.35 \\
& + \beta_{\text{OnlineSecurityYes:TechSupportYes}}0.47
\end{aligned}
$$

From the effects plots and the betas we can draw the following conclusions:

- Senior Citizen are more likely to Churn than non Senior Citizens as the have an odds of 1.3 against non senior citizens.
- We can understand tenure very easily thanks to the plot of effects. We can see how old clients of the company (old in terms of months in the company) are less probable to live the company although it smoothes this behaviour as the client reaches 40 months, this is a very important variable in order to explain churns.
- Those clients who don't have a phone service have an odds of 3.2 for leaving compared to those who only have 1 line.
- The clients who have Online Backup are less likely to live the company, with an odds of 0.72 compared to those who have not.
- Monthly charges has a linear relation with the probability of Churn, in other words, the probability of leaving is higher as the Montly Charges become higher. The odds of leaving for every unit of Monthly Charges is 1.0344. This is also a very important variable in our model.
- Those clients which have a shorter contract effect are more prone to leave than the others. We can see how the odds of leaving for Two year effect contracts is 0.11 compared to Month-to-Month. So the probability of leaving for those who have month-to-month contract are 9 times higher.
- Paperless Billing has also an effect with an odds ratio of 1.42 of yes against no.
- Lastly we can check the effect of the interaction between Online Security and Tech Support. If the client has Tech Support will be less likely to leave, otherwise will be more likely to leave, especially if she/he has not Online Security either.

Residual analysis:

We can see the effects on having an unbalanced dataset in our residual/Goodness of fit analysis. We can interpret from the plots that our residuals are far more likely to have extreme positive values rather than negative ones. In fact they are very related to the conclusions of the model interpretation. As the combination of variables gets more prone to not churn we will see more influential values in the positive axis.

Standarize test

```
test_new$OnlineBackup     <- test_new$OnlineBackup     %>% as.character()
test_new$OnlineSecurity     <- test_new$OnlineSecurity     %>% as.character()
test_new$DeviceProtection<- test_new$DeviceProtection %>% as.character()
test_new$TechSupport      <- test_new$TechSupport      %>% as.character()
test_new$StreamingTV      <- test_new$StreamingTV      %>% as.character()
test_new$StreamingMovies <- test_new$StreamingMovies %>% as.character()

test_new$OnlineBackup     <- ifelse(test_new$OnlineBackup == 'No internet
service', 'No', test_new$OnlineBackup)
test_new$OnlineSecurity     <- ifelse(test_new$OnlineSecurity == 'No internet
service', 'No', test_new$OnlineSecurity)
test_new$DeviceProtection <- ifelse(test_new$DeviceProtection == 'No internet
service', 'No', test_new$DeviceProtection)
test_new$TechSupport      <- ifelse(test_new$TechSupport == 'No internet
service', 'No', test_new$TechSupport)
test_new$StreamingTV      <- ifelse(test_new$StreamingTV == 'No internet
service', 'No', test_new$StreamingTV)
test_new$StreamingMovies  <- ifelse(test_new$StreamingMovies == 'No internet
service', 'No', test_new$StreamingTV)

test_new$OnlineBackup     <- test_new$OnlineBackup     %>% as.factor()
test_new$OnlineSecurity     <- test_new$OnlineSecurity     %>% as.factor()
test_new$DeviceProtection<- test_new$DeviceProtection %>% as.factor()
test_new$TechSupport      <- test_new$TechSupport      %>% as.factor()
test_new$StreamingTV      <- test_new$StreamingTV      %>% as.factor()
test_new$StreamingMovies <- test_new$StreamingMovies %>% as.factor()
```

Goodness of fit

```
final_model <- mod7
dim(test_new)

## [1] 1761   23

predictions <- predict(final_model,test_new, type="response")
test_new$PredictedChurn <- ifelse(predictions > 0.5,"Yes","No") %>% as.factor
val <- table(test_new$PredictedChurn, test_new$Churn)

val %>% knitr::kable()

accuracy <- sum(diag(val))/sum(val)
TP <- val[2,2]
FN <- val[1,2]
FP <- val[2,1]

accuracy <- sum(diag(val))/sum(val)
Recall <- TP/(TP+FN)
Precision <- TP / (TP + FP)
F1 <- 2 * (Precision * Recall) / (Precision + Recall)
```

```
GOF <- rbind(accuracy, Recall, Precision, F1)
colnames(GOF) <- "Metrics"
GOF %>% round(2) %>% knitr::kable()
```

Confusion Matrix:

	No	Yes
No	1210	189
Yes	113	249

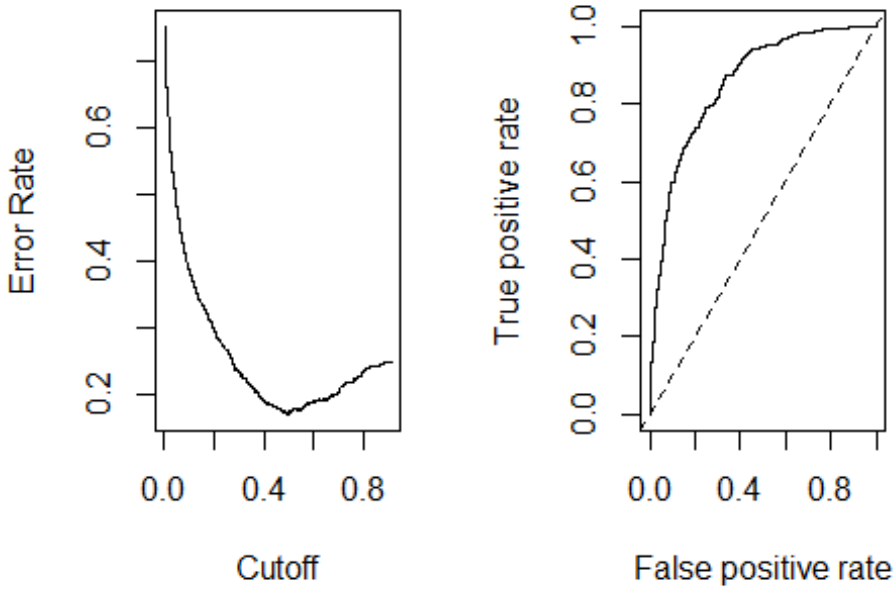
Metrics:

	Metrics
accuracy	0.83
Recall	0.57
Precision	0.69
F1	0.62

```
library("ROCR")
```

```
## Warning: package 'ROCR' was built under R version 4.3.2
```

```
dadesroc<-prediction(predict(final_model, newdata =
test_new,type="response"),test_new$Churn)
par(mfrow=c(1,2))
plot(performance(dadesroc,"err"))
plot(performance(dadesroc,"tpr","fpr"))
abline(0,1,lty=2)
```




```
predictions <- predict(final_model,test_new, type="response")
test_new$PredictedChurn <- ifelse(predictions > 0.4,"Yes","No") %>% as.factor
val <- table(test_new$PredictedChurn, test_new$Churn)

table(test_new$PredictedChurn)

##
##   No  Yes
## 1263  498

table(test_new$Churn)

##
##   No  Yes
## 1323  438

TP <- val[2,2]
FN <- val[1,2]
FP <- val[2,1]

accuracy <- sum(diag(val))/sum(val)
Recall <- TP/(TP+FN)
Precision <- TP / (TP + FP)
F1 <- 2 * (Precision * Recall) / (Precision + Recall)

GOF <- rbind(accuracy, Recall, Precision, F1)
```

```r
colnames(GOF) <- "Metrics"
GOF %>% round(2) %>% knitr::kable()
```

	Metrics
accuracy	0.81
Recall	0.68
Precision	0.60
F1	0.64

We utilized 20% of our data to establish the goodness of fit, applying the final model to predict churn within our test set. Subsequently, we applied a 0.5 threshold, obtaining an accuracy of 0.83. Notwithstanding the semblance of a good fit, misleading interpretations could arise due to an imbalanced dataset.

Exploration of the metrics table highlights a F1 score of 0.62 and a recall of 0.57, providing a clearer picture. It led us to the decision of changing the threshold to 0.4 based on the Receiver Operating Characteristic (ROC) curve. Despite a minor decrease in accuracy to 0.81, we noted an improvement in the F1 score to 0.64 and a significant increase in recall (0.68). We consider this shift important because the company is more concerned about false positives since they have a greater impact on the business than false negatives.