Customer\_Churn

Chirag Ahluwalia

3/24/2020

#### Read the Data

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.0 ✓ purrr 0.3.3  
## ✓ tibble 2.1.3 ✓ dplyr 0.8.5  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

#library(lubridate)  
library(funModeling)

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## funModeling v.1.9.3 :)  
## Examples and tutorials at livebook.datascienceheroes.com  
## / Now in Spanish: librovivodecienciadedatos.ai

#### Read the Data

library(readxl) # fast excel reader  
churn <- read\_excel("Churn.xlsx", sheet = "Case Data")

#### Show the raw data

#library(knitr) # the package that renders R markdown and has some good additional functionality  
#kable(churn)  
glimpse(churn)

## Observations: 6,347  
## Variables: 13  
## $ ID <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13…  
## $ `Customer Age (in months)` <dbl> 67, 67, 55, 63, 57, 58, 57, 46, 56, 56, 5…  
## $ `Churn (1 = Yes, 0 = No)` <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ `CHI Score Month 0` <dbl> 0, 62, 0, 231, 43, 138, 180, 116, 78, 78,…  
## $ `CHI Score 0-1` <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, -37, -1…  
## $ `Support Cases Month 0` <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,…  
## $ `Support Cases 0-1` <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0, 0, 0, …  
## $ `SP Month 0` <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0, 0,…  
## $ `SP 0-1` <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0,…  
## $ `Logins 0-1` <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7, 14, 0…  
## $ `Blog Articles 0-1` <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3, 0, 9, …  
## $ `Views 0-1` <dbl> 0, -16, 0, 21996, 9, -33, 907, 38, 0, 30,…  
## $ `Days Since Last Login 0-1` <dbl> 31, 31, 31, 0, 31, 0, 0, 6, 7, 14, 0, 31,…

#### Checking Missing Values

dim(churn)

## [1] 6347 13

str(churn)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 6347 obs. of 13 variables:  
## $ ID : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Customer Age (in months) : num 67 67 55 63 57 58 57 46 56 56 ...  
## $ Churn (1 = Yes, 0 = No) : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CHI Score Month 0 : num 0 62 0 231 43 138 180 116 78 78 ...  
## $ CHI Score 0-1 : num 0 4 0 1 -1 -10 -5 -11 -7 -37 ...  
## $ Support Cases Month 0 : num 0 0 0 1 0 0 1 0 1 0 ...  
## $ Support Cases 0-1 : num 0 0 0 -1 0 0 1 0 -2 0 ...  
## $ SP Month 0 : num 0 0 0 3 0 0 3 0 3 0 ...  
## $ SP 0-1 : num 0 0 0 0 0 0 3 0 0 0 ...  
## $ Logins 0-1 : num 0 0 0 167 0 43 13 0 -9 -7 ...  
## $ Blog Articles 0-1 : num 0 0 0 -8 0 0 -1 0 1 0 ...  
## $ Views 0-1 : num 0 -16 0 21996 9 ...  
## $ Days Since Last Login 0-1: num 31 31 31 0 31 0 0 6 7 14 ...

summary(churn)

## ID Customer Age (in months) Churn (1 = Yes, 0 = No)  
## Min. : 1 Min. : 0.0 Min. :0.00000   
## 1st Qu.:1588 1st Qu.: 5.0 1st Qu.:0.00000   
## Median :3174 Median :11.0 Median :0.00000   
## Mean :3174 Mean :13.9 Mean :0.05089   
## 3rd Qu.:4760 3rd Qu.:20.0 3rd Qu.:0.00000   
## Max. :6347 Max. :67.0 Max. :1.00000   
## CHI Score Month 0 CHI Score 0-1 Support Cases Month 0  
## Min. : 0.00 Min. :-125.000 Min. : 0.0000   
## 1st Qu.: 24.50 1st Qu.: -8.000 1st Qu.: 0.0000   
## Median : 87.00 Median : 0.000 Median : 0.0000   
## Mean : 87.32 Mean : 5.059 Mean : 0.7063   
## 3rd Qu.:139.00 3rd Qu.: 15.000 3rd Qu.: 1.0000   
## Max. :298.00 Max. : 208.000 Max. :32.0000   
## Support Cases 0-1 SP Month 0 SP 0-1 Logins 0-1   
## Min. :-29.000000 Min. :0.0000 Min. :-4.00000 Min. :-293.00   
## 1st Qu.: 0.000000 1st Qu.:0.0000 1st Qu.: 0.00000 1st Qu.: -1.00   
## Median : 0.000000 Median :0.0000 Median : 0.00000 Median : 2.00   
## Mean : -0.006932 Mean :0.8128 Mean : 0.03017 Mean : 15.73   
## 3rd Qu.: 0.000000 3rd Qu.:2.6667 3rd Qu.: 0.00000 3rd Qu.: 23.00   
## Max. : 31.000000 Max. :4.0000 Max. : 4.00000 Max. : 865.00   
## Blog Articles 0-1 Views 0-1 Days Since Last Login 0-1  
## Min. :-75.0000 Min. :-28322.00 Min. :-648.000   
## 1st Qu.: 0.0000 1st Qu.: -11.00 1st Qu.: 0.000   
## Median : 0.0000 Median : 0.00 Median : 0.000   
## Mean : 0.1572 Mean : 96.31 Mean : 1.765   
## 3rd Qu.: 0.0000 3rd Qu.: 27.00 3rd Qu.: 3.000   
## Max. :217.0000 Max. :230414.00 Max. : 61.000

# Checking Missing Values   
library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

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

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

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

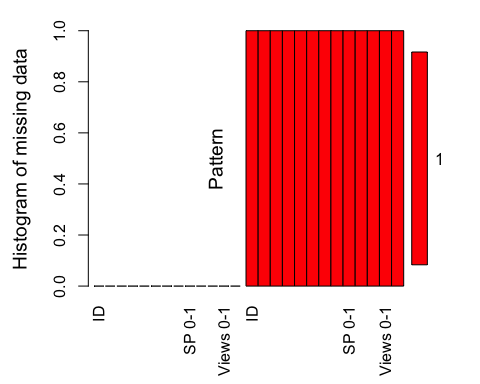
## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

aggr\_plot = aggr(churn,   
 col = c("red", "blue"),   
 numbers = TRUE,   
 prop = TRUE,   
 sortVars = TRUE,   
 labels = names(churn),   
 cex.axis = 1,   
 gap = 0,   
 ylab = c("Histogram of missing data", "Pattern"))



##   
## Variables sorted by number of missings:   
## Variable Count  
## ID 0  
## Customer Age (in months) 0  
## Churn (1 = Yes, 0 = No) 0  
## CHI Score Month 0 0  
## CHI Score 0-1 0  
## Support Cases Month 0 0  
## Support Cases 0-1 0  
## SP Month 0 0  
## SP 0-1 0  
## Logins 0-1 0  
## Blog Articles 0-1 0  
## Views 0-1 0  
## Days Since Last Login 0-1 0

# This shows that there are no missing data in our Dataset, Good Start!

#### Handling the Age variables for the initial Hypothesis

churn$`Customer Age (in months)` <- factor(churn$`Customer Age (in months)`)  
levels(churn$`Customer Age (in months)`) = list("Less than 6 Months" = c(0:5),  
 "Btw 6 to 14 Months" = c(6:14),   
 "More than 14 Months" = c(15:67))  
library(Hmisc)  
describe(churn$`Customer Age (in months)`)

## churn$`Customer Age (in months)`   
## n missing distinct   
## 6347 0 3   
##   
## Value Less than 6 Months Btw 6 to 14 Months More than 14 Months  
## Frequency 1669 2284 2394  
## Proportion 0.263 0.360 0.377

#### Handling the Dependent Variable

# converting all the variables into factor which have less than 4 unique values   
  
col\_names <- sapply(churn, function(col) length(unique(col)) < 4)  
churn[ , col\_names] <- lapply(churn[ , col\_names] , factor)  
  
  
library(Hmisc)  
describe(churn$`Churn (1 = Yes, 0 = No)`)

## churn$`Churn (1 = Yes, 0 = No)`   
## n missing distinct   
## 6347 0 2   
##   
## Value 0 1  
## Frequency 6024 323  
## Proportion 0.949 0.051

## Question1

Let us check Wall’s belief about the dependence of churn rates on customer age supported by the data? Let us try to do it through Vistualisation

png(filename= "Age Over Churn Rate.png",   
 width = 888,  
 height = 571)  
  
ggplot(data = churn, mapping = aes(x =`Customer Age (in months)`)) +  
geom\_bar(color="black",fill="orange")+  
facet\_wrap(~ `Churn (1 = Yes, 0 = No)`)+  
ggtitle("Vistualising Age Categories over Churn")+  
 xlab("Customer Age - in months")+  
 theme\_bw()  
  
xtabs(~`Churn (1 = Yes, 0 = No)` + `Customer Age (in months)`, data = churn)

## Customer Age (in months)  
## Churn (1 = Yes, 0 = No) Less than 6 Months Btw 6 to 14 Months  
## 0 1643 2119  
## 1 26 165  
## Customer Age (in months)  
## Churn (1 = Yes, 0 = No) More than 14 Months  
## 0 2262  
## 1 132

library(Hmisc)  
describe(churn$`Customer Age (in months)`)

## churn$`Customer Age (in months)`   
## n missing distinct   
## 6347 0 3   
##   
## Value Less than 6 Months Btw 6 to 14 Months More than 14 Months  
## Frequency 1669 2284 2394  
## Proportion 0.263 0.360 0.377

# Churn - Yes(1) - means that the Customer left the application and No(0) - means that the customer is still using the application   
  
# Now we see that wall's belief is correct for people who are still using the application;   
# People who have been using the application for more than 14 months are still likely to stay with us   
  
# However, People who might leave us are from the age range of 6 to 14 months, which was the correct assessment of Walls. However, people who have been using the system for more than 14 months are leaving more than that people who recently started using the system.   
  
# This is a potential problem, as it negates our Hypothesis

## Let us check other variables

## Question 2

Let us build a logistic regression Model

#install.packages("aod")  
library(aod)

##   
## Attaching package: 'aod'

## The following object is masked from 'package:survival':  
##   
## rats

mylogit <- glm( `Churn (1 = Yes, 0 = No)`~., data = churn, family = "binomial")  
summary(mylogit)

##   
## Call:  
## glm(formula = `Churn (1 = Yes, 0 = No)` ~ ., family = "binomial",   
## data = churn)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9182 -0.3558 -0.2463 -0.1859 3.4687   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -2.591e+00 2.956e-01 -8.767  
## ID -2.365e-04 4.262e-05 -5.549  
## `Customer Age (in months)`Btw 6 to 14 Months 1.593e+00 2.502e-01 6.367  
## `Customer Age (in months)`More than 14 Months 1.071e+00 2.747e-01 3.899  
## `CHI Score Month 0` -1.023e-02 1.282e-03 -7.980  
## `CHI Score 0-1` -4.081e-03 2.546e-03 -1.603  
## `Support Cases Month 0` -7.521e-02 1.003e-01 -0.749  
## `Support Cases 0-1` 1.178e-01 8.853e-02 1.330  
## `SP Month 0` -2.830e-02 1.004e-01 -0.282  
## `SP 0-1` -8.635e-03 7.709e-02 -0.112  
## `Logins 0-1` 7.771e-04 2.167e-03 0.359  
## `Blog Articles 0-1` 4.448e-04 2.267e-02 0.020  
## `Views 0-1` -9.791e-05 4.230e-05 -2.315  
## `Days Since Last Login 0-1` 6.331e-03 3.177e-03 1.993  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## ID 2.87e-08 \*\*\*  
## `Customer Age (in months)`Btw 6 to 14 Months 1.92e-10 \*\*\*  
## `Customer Age (in months)`More than 14 Months 9.65e-05 \*\*\*  
## `CHI Score Month 0` 1.47e-15 \*\*\*  
## `CHI Score 0-1` 0.1090   
## `Support Cases Month 0` 0.4536   
## `Support Cases 0-1` 0.1834   
## `SP Month 0` 0.7781   
## `SP 0-1` 0.9108   
## `Logins 0-1` 0.7199   
## `Blog Articles 0-1` 0.9843   
## `Views 0-1` 0.0206 \*   
## `Days Since Last Login 0-1` 0.0463 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2553.1 on 6346 degrees of freedom  
## Residual deviance: 2298.1 on 6333 degrees of freedom  
## AIC: 2326.1  
##   
## Number of Fisher Scoring iterations: 7

exp(coef(mylogit))

## (Intercept)   
## 0.07494508   
## ID   
## 0.99976355   
## `Customer Age (in months)`Btw 6 to 14 Months   
## 4.92060070   
## `Customer Age (in months)`More than 14 Months   
## 2.91881824   
## `CHI Score Month 0`   
## 0.98982181   
## `CHI Score 0-1`   
## 0.99592741   
## `Support Cases Month 0`   
## 0.92755171   
## `Support Cases 0-1`   
## 1.12498864   
## `SP Month 0`   
## 0.97209879   
## `SP 0-1`   
## 0.99140245   
## `Logins 0-1`   
## 1.00077740   
## `Blog Articles 0-1`   
## 1.00044490   
## `Views 0-1`   
## 0.99990210   
## `Days Since Last Login 0-1`   
## 1.00635153

exp(cbind(OR = coef(mylogit), confint(mylogit)))

## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## OR 2.5 % 97.5 %  
## (Intercept) 0.07494508 0.04133068 0.1318496  
## ID 0.99976355 0.99968038 0.9998475  
## `Customer Age (in months)`Btw 6 to 14 Months 4.92060070 3.05221350 8.1645041  
## `Customer Age (in months)`More than 14 Months 2.91881824 1.72444389 5.0735685  
## `CHI Score Month 0` 0.98982181 0.98731816 0.9922952  
## `CHI Score 0-1` 0.99592741 0.99099807 1.0009389  
## `Support Cases Month 0` 0.92755171 0.74290588 1.1014597  
## `Support Cases 0-1` 1.12498864 0.96681251 1.3675430  
## `SP Month 0` 0.97209879 0.79905582 1.1852611  
## `SP 0-1` 0.99140245 0.85196114 1.1529652  
## `Logins 0-1` 1.00077740 0.99615153 1.0045225  
## `Blog Articles 0-1` 1.00044490 0.95174516 1.0292632  
## `Views 0-1` 0.99990210 0.99982227 0.9999951  
## `Days Since Last Login 0-1` 1.00635153 1.00044524 1.0129888