01\_EDA

Matt Dube

12/4/2018

Table of Contents

Load libraries

library(here)  
library(readr)  
library(dplyr)  
library(skimr)  
library(broom)

Load data

customer <- read\_csv(here("00\_Data/raw", "WA\_Fn-UseC\_-Telco-Customer-Churn.csv"))

Quick check of front/back/structure

customer %>% head(10)

## # A tibble: 10 x 21  
## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## <chr> <chr> <dbl> <chr> <chr> <dbl> <chr>   
## 1 7590-VHVEG Female 0 Yes No 1 No   
## 2 5575-GNVDE Male 0 No No 34 Yes   
## 3 3668-QPYBK Male 0 No No 2 Yes   
## 4 7795-CFOCW Male 0 No No 45 No   
## 5 9237-HQITU Female 0 No No 2 Yes   
## 6 9305-CDSKC Female 0 No No 8 Yes   
## 7 1452-KIOVK Male 0 No Yes 22 Yes   
## 8 6713-OKOMC Female 0 No No 10 No   
## 9 7892-POOKP Female 0 Yes No 28 Yes   
## 10 6388-TABGU Male 0 No Yes 62 Yes   
## # ... with 14 more variables: MultipleLines <chr>, InternetService <chr>,  
## # OnlineSecurity <chr>, OnlineBackup <chr>, DeviceProtection <chr>,  
## # TechSupport <chr>, StreamingTV <chr>, StreamingMovies <chr>,  
## # Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,  
## # MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <chr>

customer %>% tail(10)

## # A tibble: 10 x 21  
## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## <chr> <chr> <dbl> <chr> <chr> <dbl> <chr>   
## 1 9767-FFLEM Male 0 No No 38 Yes   
## 2 0639-TSIQW Female 0 No No 67 Yes   
## 3 8456-QDAVC Male 0 No No 19 Yes   
## 4 7750-EYXWZ Female 0 No No 12 No   
## 5 2569-WGERO Female 0 No No 72 Yes   
## 6 6840-RESVB Male 0 Yes Yes 24 Yes   
## 7 2234-XADUH Female 0 Yes Yes 72 Yes   
## 8 4801-JZAZL Female 0 Yes Yes 11 No   
## 9 8361-LTMKD Male 1 Yes No 4 Yes   
## 10 3186-AJIEK Male 0 No No 66 Yes   
## # ... with 14 more variables: MultipleLines <chr>, InternetService <chr>,  
## # OnlineSecurity <chr>, OnlineBackup <chr>, DeviceProtection <chr>,  
## # TechSupport <chr>, StreamingTV <chr>, StreamingMovies <chr>,  
## # Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,  
## # MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <chr>

skim(customer)

## Skim summary statistics  
## n obs: 7043   
## n variables: 21   
##   
## -- Variable type:character -------------------------------------------------------------------------------------  
## variable missing complete n min max empty n\_unique  
## Churn 0 7043 7043 2 3 0 2  
## Contract 0 7043 7043 8 14 0 3  
## customerID 0 7043 7043 10 10 0 7043  
## Dependents 0 7043 7043 2 3 0 2  
## DeviceProtection 0 7043 7043 2 19 0 3  
## gender 0 7043 7043 4 6 0 2  
## InternetService 0 7043 7043 2 11 0 3  
## MultipleLines 0 7043 7043 2 16 0 3  
## OnlineBackup 0 7043 7043 2 19 0 3  
## OnlineSecurity 0 7043 7043 2 19 0 3  
## PaperlessBilling 0 7043 7043 2 3 0 2  
## Partner 0 7043 7043 2 3 0 2  
## PaymentMethod 0 7043 7043 12 25 0 4  
## PhoneService 0 7043 7043 2 3 0 2  
## StreamingMovies 0 7043 7043 2 19 0 3  
## StreamingTV 0 7043 7043 2 19 0 3  
## TechSupport 0 7043 7043 2 19 0 3  
##   
## -- Variable type:numeric ---------------------------------------------------------------------------------------  
## variable missing complete n mean sd p0 p25 p50  
## MonthlyCharges 0 7043 7043 64.76 30.09 18.25 35.5 70.35  
## SeniorCitizen 0 7043 7043 0.16 0.37 0 0 0   
## tenure 0 7043 7043 32.37 24.56 0 9 29   
## TotalCharges 11 7032 7043 2283.3 2266.77 18.8 401.45 1397.47  
## p75 p100 hist  
## 89.85 118.75 <U+2587><U+2581><U+2583><U+2582><U+2586><U+2585><U+2585><U+2582>  
## 0 1 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2582>  
## 55 72 <U+2587><U+2583><U+2583><U+2582><U+2582><U+2583><U+2583><U+2585>  
## 3794.74 8684.8 <U+2587><U+2583><U+2582><U+2582><U+2581><U+2581><U+2581><U+2581>

It looks like there are no missing values except in TotalCharges, let’s do an explicit check by column to be sure.

customer %>%   
 select(everything()) %>%   
 summarise\_all(funs(sum(is.na(.))))

## # A tibble: 1 x 21  
## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## <int> <int> <int> <int> <int> <int> <int>  
## 1 0 0 0 0 0 0 0  
## # ... with 14 more variables: MultipleLines <int>, InternetService <int>,  
## # OnlineSecurity <int>, OnlineBackup <int>, DeviceProtection <int>,  
## # TechSupport <int>, StreamingTV <int>, StreamingMovies <int>,  
## # Contract <int>, PaperlessBilling <int>, PaymentMethod <int>,  
## # MonthlyCharges <int>, TotalCharges <int>, Churn <int>

No other missing values. Only .15% of values are missing - there is not really a wrong way to handle this.  
Here are the choices:

* delete the records - again, only 11 of 7043 records will be removed.
* create an indicator variable to show missingness - not ideal, will be 7032 0’s, and only 11 1’s.
* impute during model training - using knnImpute or bagImpute. Might consider this for the practice, not for any expected bump in model performance.

Target variable is ‘Churn’. Review class distribution.

table(customer$Churn)

##   
## No Yes   
## 5174 1869

prop.table(table(customer$Churn))

##   
## No Yes   
## 0.7346301 0.2653699

Class is 3/4 ‘No’, 1/4 ‘Yes’. Make a note to review this during modeling, may need to consider trying a couple of sampling methods to balance the classes for model fitting. Problem is binary classification, which will provide us with a number of options to try when fitting different models.

Additional notes to review during data cleaning:

* there are several binary features that can be encoded as 1/0.
  + Gender, SeniorCitizen, Partner, PhoneService
* there are several more features that have only a few classes, so are good candidates for one-hot encoding (this can be done during preProcessing, or we can let caret take care of it under the hood during training).
  + Looks like most of the other features fall into this area.
* customerID can be dropped - it’s unique, only adds ‘noise’.
* looks like only MonthlyCharges and TotalCharges are numeric. Check their correlation to determine if they are both needed.
* tenure is an integer - looks to be measured in months.