Telco Customer Churn Classification

David Schultheiss

3/4/2021

```r
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v dplyr   1.0.2
## v tibble  3.0.4     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v purrr   0.3.4
```

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

```r
library(GGally)
```

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

```r
library(car)
```

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

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

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

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

```r
library(cowplot)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

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

```r
library(class)
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

```r
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

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

Introduction

This analysis focuses on predicting customer retention using the "Telco Customer Churn" dataset, published on Kaggle. The dataset contains 21 variables. These include a customer ID variable, 19 predictors, and our target variable customer churn. Using these predictors, we can identify important variables and predict which customers are likely to leave the platform. In this paper, I will accomplish this by applying logisitic regression, KNN classification, and random forest.

Reading in the data:

```r
churn <-
  read_csv("/Users/davidschultheiss/Downloads/WA_Fn-UseC_-Telco-Customer-Churn.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   SeniorCitizen = col_double(),
##   tenure = col_double(),
##   MonthlyCharges = col_double(),
##   TotalCharges = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```r
sum(!complete.cases(churn))
```

```
## [1] 11
```

```r
churn = churn[complete.cases(churn), ]
churn$SeniorCitizen = as.factor(ifelse(churn$SeniorCitizen==1, 'Yes', 'No'))
glimpse(churn)
```

```
## Rows: 7,032
## Columns: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CF...
## $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Femal...
## $ SeniorCitizen    <fct> No, No, No, No, No, No, No, No, No, No, No, No, No...
## $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "...
## $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "...
## $ tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49...
## $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No...
## $ MultipleLines    <chr> "No phone service", "No", "No", "No phone service"...
## $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber ...
## $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes"...
## $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No",...
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No",...
## $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "...
## $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", ...
## $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "...
## $ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "O...
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No...
## $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check"...
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 2...
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1...
## $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No",...
```

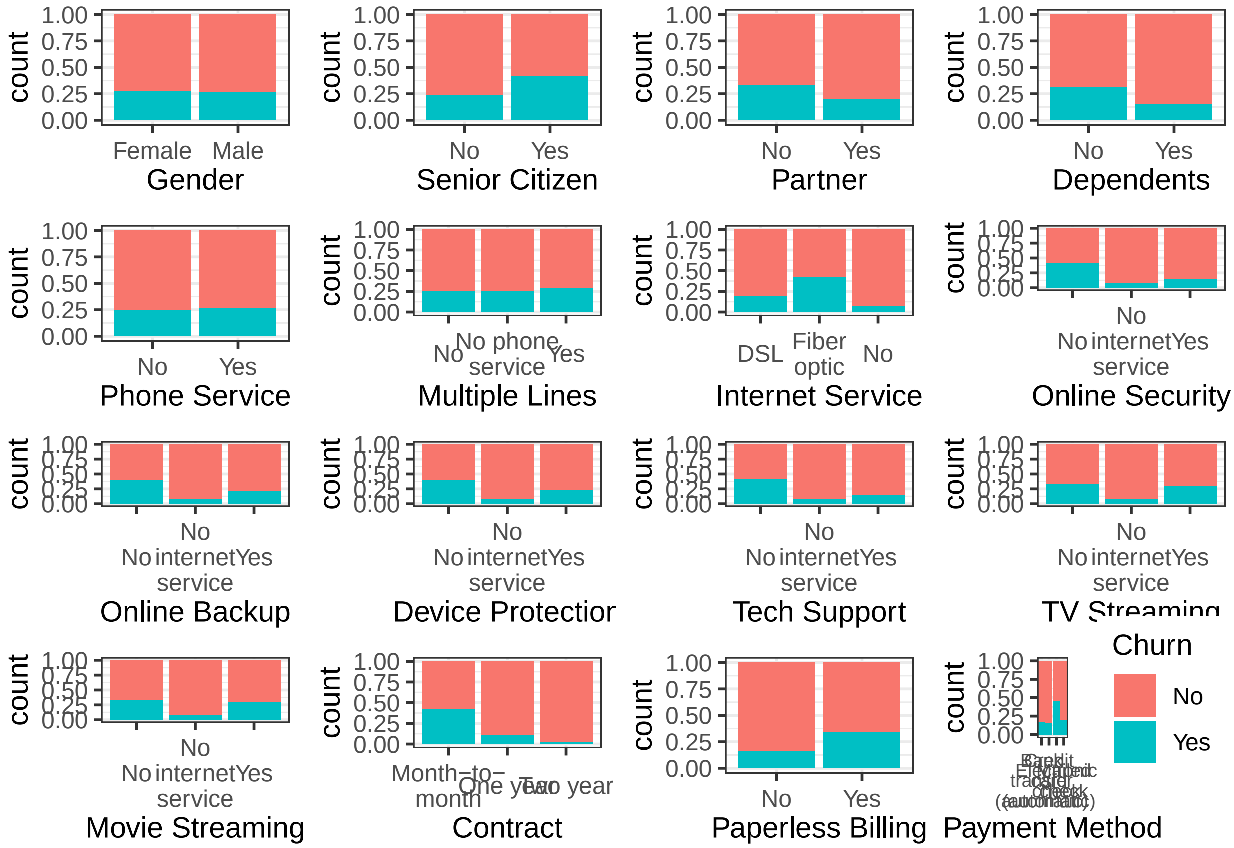
Visualization

First, a look into the categorical variables.

```
ggtheme = theme_bw()+
  theme(axis.text.x= element_text(angle= 0, hjust= .5, vjust= .5),
        legend.position= "none")

ggfun1 = function(aesx) {
  ggplot(churn, aes(x= aesx, fill= Churn)) +
    geom_bar(position='fill') +
    ggtheme +
    scale_x_discrete(labels = function(x) str_wrap(x, width = 10))}

plot_grid(
ggfun1(churn$gender) +
  xlab('Gender'),
ggfun1(churn$SeniorCitizen) +
  xlab('Senior Citizen'),
ggfun1(churn$Partner) +
  xlab('Partner'),
ggfun1(churn$Dependents) +
  xlab('Dependents'),
ggfun1(churn$PhoneService) +
  xlab('Phone Service'),
ggfun1(churn$MultipleLines) +
  xlab('Multiple Lines'),
ggfun1(churn$InternetService) +
  xlab('Internet Service'),
ggfun1(churn$OnlineSecurity) +
  xlab('Online Security'),
ggfun1(churn$OnlineBackup) +
  xlab('Online Backup'),
ggfun1(churn$DeviceProtection) +
  xlab('Device Protection'),
ggfun1(churn$TechSupport) +
  xlab('Tech Support'),
ggfun1(churn$StreamingTV) +
  xlab('TV Streaming'),
ggfun1(churn$StreamingMovies) +
  xlab('Movie Streaming'),
ggfun1(churn$Contract) +
  xlab('Contract'),
ggfun1(churn$PaperlessBilling) +
  xlab('Paperless Billing'),
ggfun1(churn$PaymentMethod) +
  theme(axis.text.x= element_text(size= 7), legend.position = 'right') +
  xlab('Payment Method')
)
```



The teal bar in these plots shows the proportion for each category that churned. For instance, senior citizens are more likely to churn. People with a partner and/or dependants are less likely to churn. Some other notable variables are:

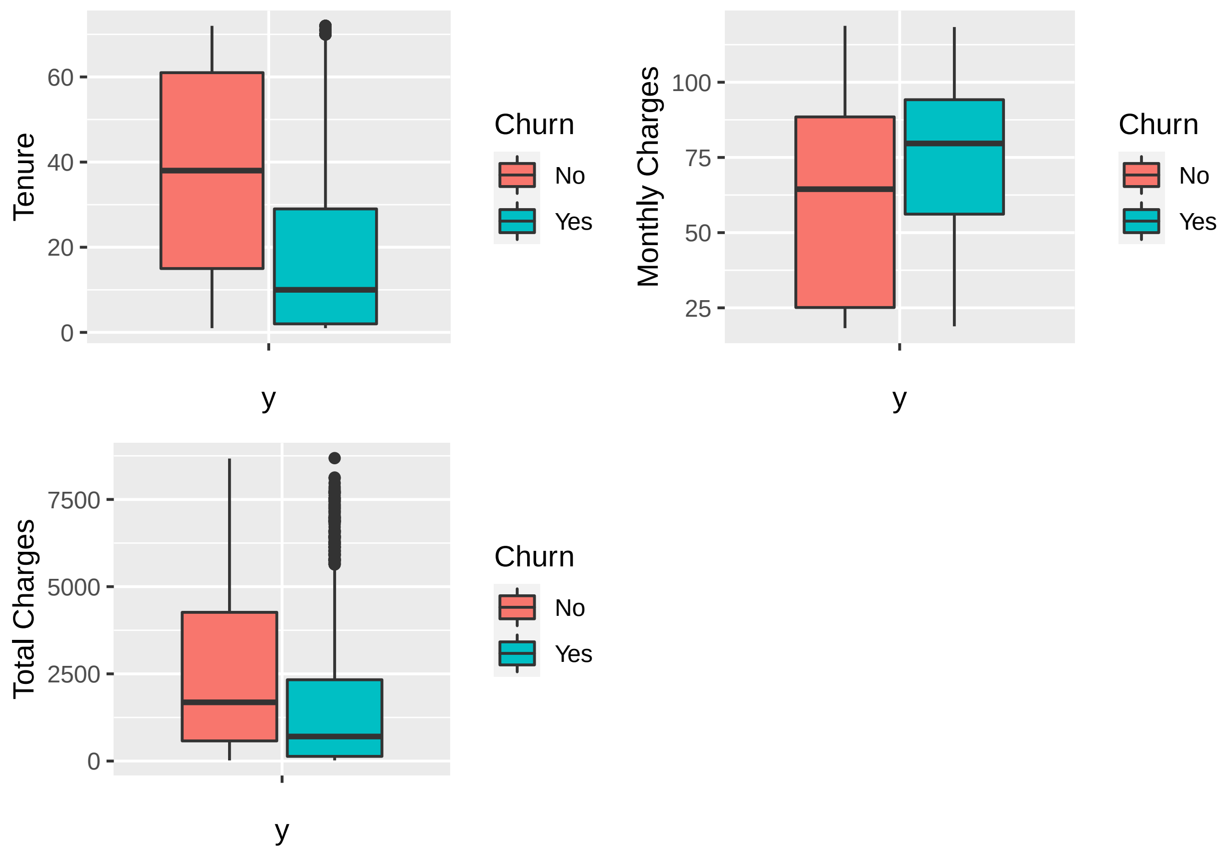
- Internet Service: Customers paying for fiber optic are more likely to churn.
- Online Security: Customers paying for online security are less likely to churn.
- Contract: Customers subscribing month-to-month are more likely to churn.
- Paperless Billing: Customers using paperless billing are more likely to churn.
- Payment Method: Customers paying with electronic check are more likely to churn.

Next we can examine our three continuous variables.

```
ggfun2 = function(aesx) {
  ggplot(churn, aes(x= aesx, y= ' ', fill= Churn)) +
    geom_boxplot() +
    coord_flip()
  }

plot_grid(
ggfun2(churn$tenure) +
  xlab('Tenure'),
ggfun2(churn$MonthlyCharges) +
  xlab('Monthly Charges'),
ggfun2(churn$TotalCharges) +
```

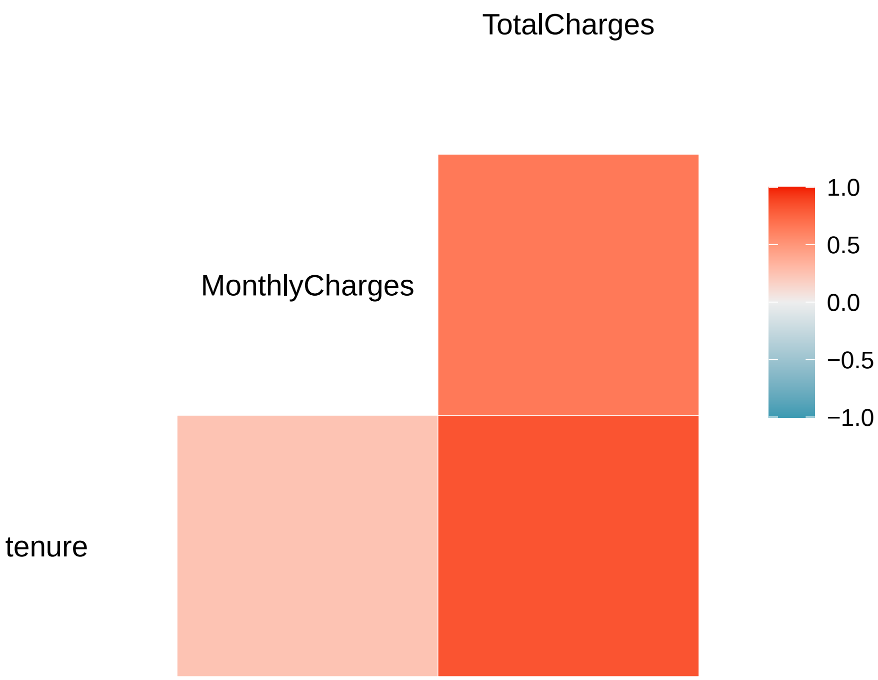
```r
  xlab('Total Charges')
)
```



Median tenure of customers who churn is about 10 months. Additonally, customers leaving the service have higher monthly charges.

```r
ggcorr(churn)
```

```
## Warning in ggcorr(churn): data in column(s) 'customerID', 'gender',
## 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines',
## 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
## 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',
## 'PaymentMethod', 'Churn' are not numeric and were ignored
```



Total charges has high positive correlation with tenure and monthly charges.

Data Processing

```
churn = churn %>%
  select(-customerID)

churn = data.frame(lapply(churn, function(x) {
  gsub("No internet service", "No", x)}))
churn = data.frame(lapply(churn, function(x) {
  gsub("No phone service", "No", x)}))

#chr -> num, scaling
int_cols = c('tenure', 'MonthlyCharges', 'TotalCharges')
churn[int_cols] = sapply(churn[int_cols], as.numeric)
churn[int_cols] = sapply(churn[int_cols], scale)

#chr -> Factor
churn[sapply(churn, is.character)] <- lapply(churn[sapply(churn, is.character)],
                                              as.factor)
#Split data into training and test sets
set.seed(1)
training = sample(1:nrow(churn), .75*nrow(churn))

train = churn[training, ]
test = churn[-training, ]
```

Modeling

Logistic Regression

```r
logit = glm(data= train, Churn~., family= 'binomial')
vif(logit)
```

```
##                        GVIF Df GVIF^(1/(2*Df))
## gender             1.004606  1        1.002300
## SeniorCitizen      1.130070  1        1.063047
## Partner            1.378809  1        1.174227
## Dependents         1.293785  1        1.137447
## tenure            15.871746  1        3.983936
## PhoneService      34.847729  1        5.903197
## MultipleLines      7.370814  1        2.714924
## InternetService  386.866131  2        4.434965
## OnlineSecurity     4.917506  1        2.217545
## OnlineBackup       6.312303  1        2.512430
## DeviceProtection   6.352084  1        2.520334
## TechSupport        5.305839  1        2.303441
## StreamingTV       24.471525  1        4.946870
## StreamingMovies   24.775637  1        4.977513
## Contract           1.622529  2        1.128621
## PaperlessBilling   1.121276  1        1.058903
## PaymentMethod      1.398075  3        1.057438
## MonthlyCharges   693.713112  1       26.338434
## TotalCharges      20.511284  1        4.528939
```

Using the variance inflation factor (VIF), we can check for multicollinearity. I won't heavily emphasize pruning the model in this paper; however, based on these results I did remove the Monthly Charges variable.

```r
logit.train = train %>%
  select(-MonthlyCharges)
```