R Programming

Seyoung Jung

SDSU Statistics 610: Linear Regression Models

Introduction

In this video, we will cover:

1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Logistic Regression

GitHub Page: https://github.com/sjung-stat/Telco

Description of Our Data

Telco Customer Churn dataset

- **Data**: 7,043 rows (customers) and 21 columns (features)
- **Information**: Customers who churned last month, Demographic information about customers, etc.
- **Goal**: To predict whether the customer will churn or not
- **Source**: www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/

Preparation

- library(readr)
- library(dplyr)
- library(tidyr)
- library(caTools)
- library(ggplot2)
- library(reshape2)
- library(gridExtra)
- library(MASS)

Data Cleaning

Load the dataset into R and get a glimpse of our data.

```
telcodata <- read_csv("telcodata.csv")
glimpse(telcodata)
```

Check if there are missing data. we will delete the corresponding rows.

```
MVinfo <- apply(is.na(telcodata), 2, which)      # 11 missing data
withoutMV <- telcodata[-MVinfo$TotalCharges, ]
```

Coerce all the variables to factor variables except for customerID, SeniorCitizen, tenure, MonthlyCharges, and TotalCharges.

```
cols <- c(1, 3, 6, 19, 20)
withoutMV[,-cols] <- data.frame(apply(withoutMV[-cols], 2, as.factor))
```

Data Cleaning

Coerce the SeniorCitizen variable to a factor. 1: Yes, 0: No.

```
withoutMV <- withoutMV %>% mutate(SeniorCitizen = ifelse(SeniorCitizen == 0,
 "No", "Yes"))
withoutMV$SeniorCitizen <- as.factor(withoutMV$SeniorCitizen)
```

Get rid of the customerID variable

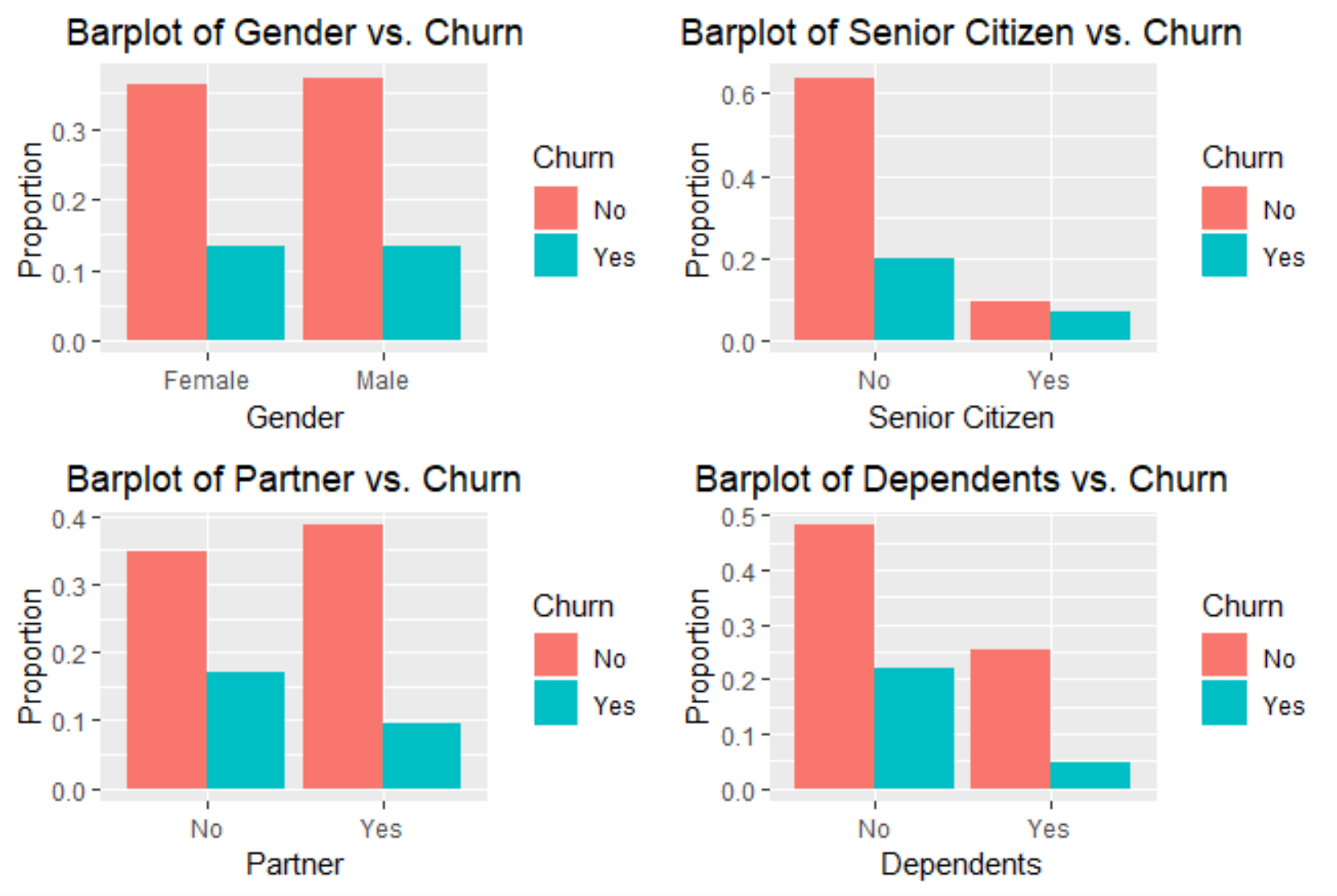
```
cleandata <- withoutMV[, 2:21]
```

Data Cleaning

Create a new column that classifies the elements by years, and get rid of tenure and TotalCharges.

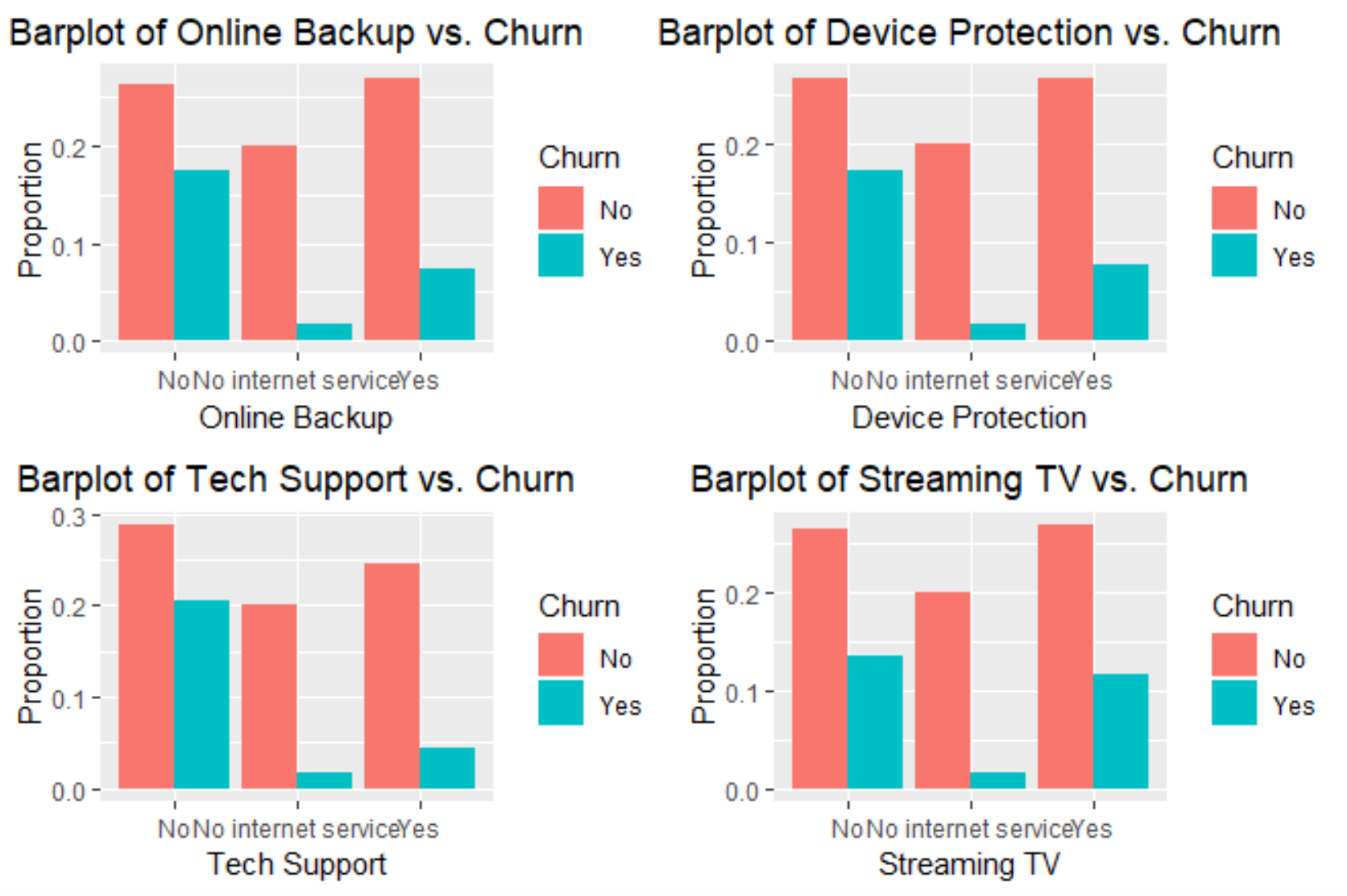
```
cleandata <- mutate(cleandata, Year=ifelse(tenure %in% 1:12, "0-1",
                                           ifelse(tenure %in% 13:24, "1-2",
                                                  ifelse(tenure %in% 25:36,
"2-3",
                                                         ifelse(tenure %in%
37:48, "3-4",
                                                                ifelse(tenur
e %in% 49:60, "4-5",
                                                                       ifels
e(tenure %in% 61:72, "5-6", "6-7")))))))
cleandata$Year <- as.factor(cleandata$Year)
cleandata$tenure <- NULL
cleandata$TotalCharges <- NULL
```

Exploratory Data Analysis

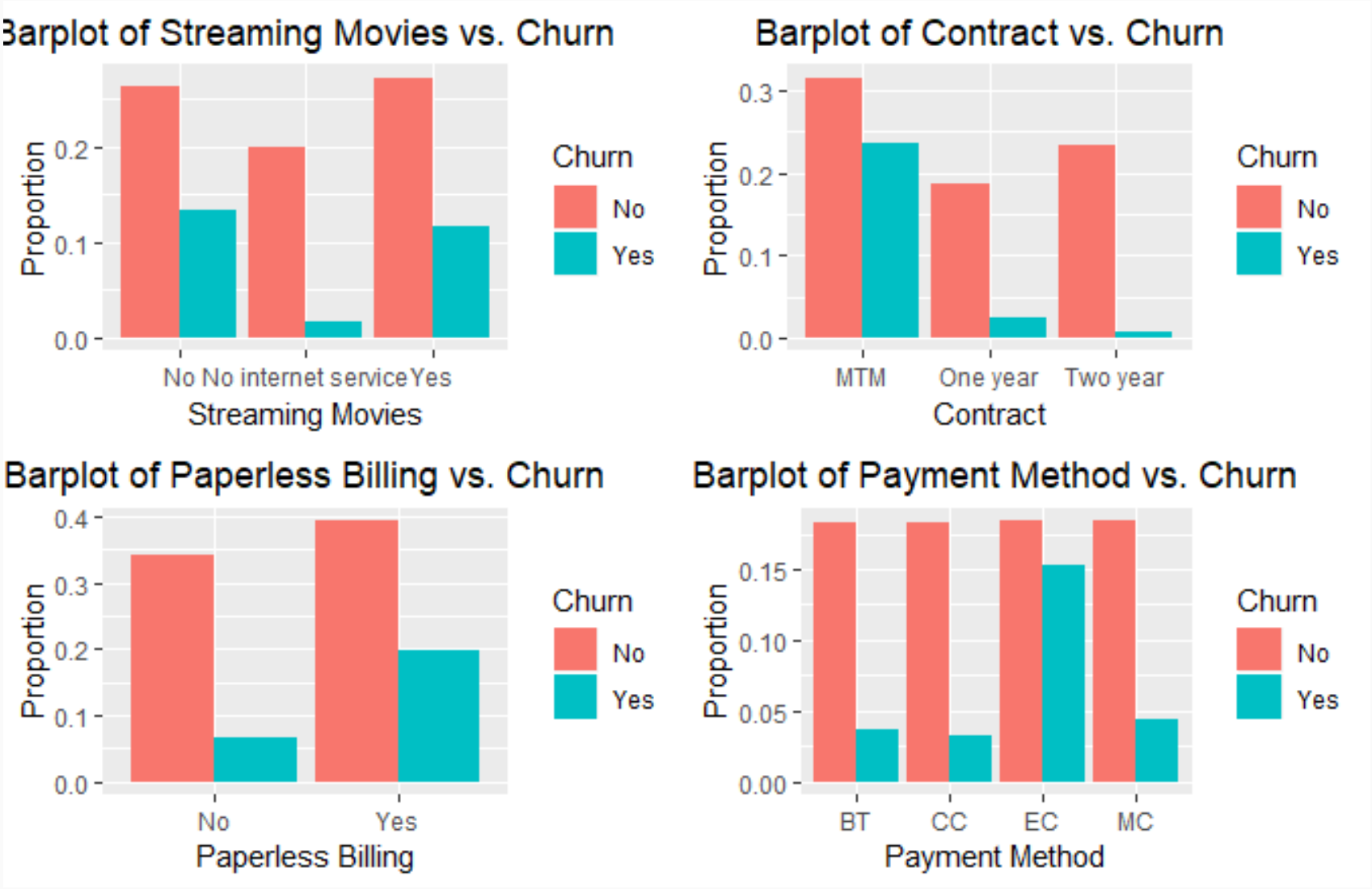


Exploratory Data Analysis

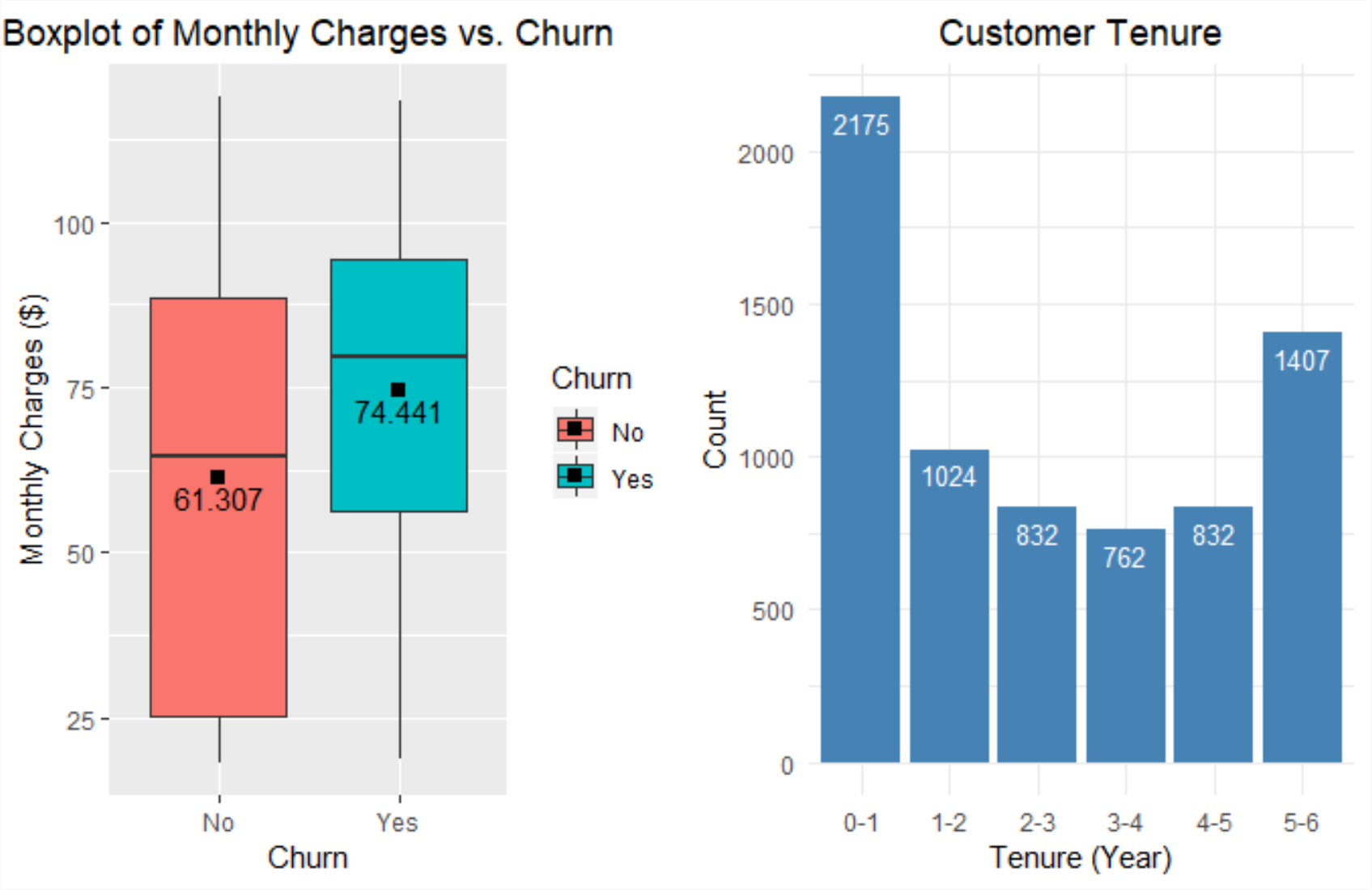
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Logistic Regression - Preparation

Make nominal categorical variables to ordinal categorical variables. And change some of our categorical variables to binary categorical variables. Regard “No Internet Service” as “No”, and change the corresponding answers accordingly.

```
cleandata<-cleandata%>%mutate(gender=ifelse(gender=="Male",1,0))
cleandata<-cleandata%>%mutate(SeniorCitizen=ifelse(SeniorCitizen=="Yes",1,0))
cleandata<-cleandata%>%mutate(Partner=ifelse(Partner=="Yes",1,0))
cleandata<-cleandata%>%mutate(Dependents=ifelse(Dependents=="Yes",1,0))
cleandata<-cleandata%>%mutate(PhoneService=ifelse(PhoneService=="Yes",1,0))
cleandata<-cleandata%>%mutate(MultipleLines=ifelse(MultipleLines=="Yes",1,0))
cleandata<-cleandata%>%mutate(OnlineSecurity=ifelse(OnlineSecurity=="Yes",1,0))
cleandata<-cleandata%>%mutate(OnlineBackup=ifelse(OnlineBackup=="Yes",1,0))
cleandata<-cleandata%>%mutate(DeviceProtection=ifelse(DeviceProtection=="Yes",1
,0))
cleandata<-cleandata%>%mutate(TechSupport=ifelse(TechSupport=="Yes",1,0))
cleandata<-cleandata%>%mutate(StreamingTV=ifelse(StreamingTV=="Yes",1,0))
cleandata<-cleandata%>%mutate(StreamingMovies=ifelse(StreamingMovies=="Yes",1,0
))
cleandata<-cleandata%>%mutate(PaperlessBilling=ifelse(PaperlessBilling=="Yes",1
,0))
cleandata<-cleandata%>%mutate(Churn=ifelse(Churn=="Yes",1,0))
```

Logistic Regression - Preparation

Standardize MonthlyCharges variable

```
cleandata$MonthlyCharges <- scale(cleandata$MonthlyCharges)
```

Creating a Baseline model

```
table(cleandata$Churn)/nrow(cleandata)   # Churn rate: 73.42%
```

```
##
##         0         1
## 0.734215 0.265785
```

Split the dataset into training and test sets

```
set.seed(12345)
sample <- sample.split(cleandata, SplitRatio=0.7)
train_data <- subset(cleandata, sample==TRUE)
test_data <- subset(cleandata, sample==FALSE)
```

Logistic Regression

Fit the logistic regression

```
log_reg1 <- glm(Churn ~ ., data=train_data, family=binomial(link="logit"))
summary(log_reg1)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0344  -0.6621  -0.2795   0.6625   3.2140
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -2.00094    1.54708  -1.293   0.1959
## gender                             -0.04226    0.07895  -0.535   0.5924
## SeniorCitizen                       0.20587    0.10153   2.028   0.0426
## Partner                            -0.03437    0.09376  -0.367   0.7140
## Dependents                         -0.24033    0.10989  -2.187   0.0287
## PhoneService                        0.42108    0.78901   0.534   0.5936
## MultipleLines                       0.49773    0.21499   2.315   0.0206
## InternetServiceFiber optic          2.05514    0.96929   2.120   0.0340
## InternetServiceNo                  -2.15347    0.98253  -2.192   0.0284
## OnlineSecurity                     -0.06996    0.21632  -0.323   0.7464
## OnlineBackup                        0.07653    0.21237   0.360   0.7186
## DeviceProtection                    0.14008    0.21444   0.653   0.5136
## TechSupport                        -0.16158    0.21897  -0.738   0.4606
## StreamingTV                         0.73557    0.39639   1.856   0.0635
## StreamingMovies                     0.73014    0.39704   1.839   0.0659
## ContractOne year                   -0.67799    0.12994  -5.218 1.81e-07
## ContractTwo year                   -1.61771    0.22582  -7.164 7.85e-13
```

Logistic Regression

```
## PaperlessBilling                        0.41703    0.09123   4.571 4.85e-06
## PaymentMethodCredit card (automatic)   -0.09860    0.13804  -0.714   0.4751
## PaymentMethodElectronic check           0.28453    0.11420   2.491   0.0127
## PaymentMethodMailed check               0.01093    0.13944   0.078   0.9375
## MonthlyCharges                         -1.37095    1.16086  -1.181   0.2376
## Year1-2                                -0.82916    0.11684  -7.097 1.28e-12
## Year2-3                                -1.26196    0.13927  -9.061  < 2e-16
## Year3-4                                -1.36110    0.15520  -8.770  < 2e-16
## Year4-5                                -1.57829    0.17383  -9.079  < 2e-16
## Year5-6                                -1.86170    0.20630  -9.024  < 2e-16
##
## (Intercept)
## gender
## SeniorCitizen                          *
## Partner
## Dependents                             *
## PhoneService
## MultipleLines                          *
## InternetServiceFiber optic             *
## InternetServiceNo                      *
## OnlineSecurity
## OnlineBackup
## DeviceProtection
## TechSupport
## StreamingTV                            .
## StreamingMovies                        .
## ContractOne year                       ***
## ContractTwo year                       ***
## PaperlessBilling                       ***
```

Logistic Regression

```
## PaymentMethodCredit card (automatic)
## PaymentMethodElectronic check        *
## PaymentMethodMailed check
## MonthlyCharges
## Year1-2                              ***
## Year2-3                              ***
## Year3-4                              ***
## Year4-5                              ***
## Year5-6                              ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5622.1  on 4810  degrees of freedom
## Residual deviance: 3962.2  on 4784  degrees of freedom
## AIC: 4016.2
##
## Number of Fisher Scoring iterations: 6
```

Logistic Regression

Check the deviance

```
anova(object=log_reg1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                4810     5622.1
## gender            1     0.88      4809     5621.2   0.34842
## SeniorCitizen     1   103.89      4808     5517.3 < 2.2e-16 ***
## Partner           1   133.36      4807     5384.0 < 2.2e-16 ***
## Dependents        1    41.21      4806     5342.8 1.368e-10 ***
## PhoneService      1     0.10      4805     5342.7   0.74941
## MultipleLines     1     6.51      4804     5336.2   0.01073 *
## InternetService   2   512.29      4802     4823.9 < 2.2e-16 ***
## OnlineSecurity    1   157.98      4801     4665.9 < 2.2e-16 ***
## OnlineBackup      1    76.54      4800     4589.3 < 2.2e-16 ***
## DeviceProtection  1    48.06      4799     4541.3 4.133e-12 ***
## TechSupport       1    93.36      4798     4447.9 < 2.2e-16 ***
## StreamingTV       1     1.96      4797     4446.0   0.16172
## StreamingMovies   1     0.39      4796     4445.6   0.52982
## Contract          2   267.96      4794     4177.6 < 2.2e-16 ***
## PaperlessBilling  1    20.10      4793     4157.5 7.340e-06 ***
## PaymentMethod     3    32.01      4790     4125.5 5.197e-07 ***
## MonthlyCharges    1     1.24      4789     4124.2   0.26550
## Year              5   162.07      4784     3962.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic Regression

Evaluate the logistic regression model

```
predict_train <- predict(log_reg1, newdata=test_data, type="response")
predict_train <- ifelse(predict_train > 0.5, 1, 0)
predict_error <- mean(predict_train != test_data$Churn)
model1 <- 1 - predict_error
print(model1)
```

```
## [1] 0.7982891
```

Logistic Regression

Variable selection

```
step <- stepAIC(log_reg1, trace=FALSE)
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Churn ~ gender + SeniorCitizen + Partner + Dependents + PhoneService +

##     MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
##     DeviceProtection + TechSupport + StreamingTV + StreamingMovies +
##     Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +
##     Year
##
## Final Model:
## Churn ~ SeniorCitizen + Dependents + PhoneService + MultipleLines +
##     InternetService + OnlineBackup + DeviceProtection + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + Year
##
##
##                Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                                     4784   3962.176 4016.176
## 2 - OnlineSecurity  1 0.1045962      4785   3962.280 4014.280
## 3         - Partner  1 0.1431430      4786   3962.423 4012.423
## 4          - gender  1 0.2789622      4787   3962.702 4010.702
## 5    - TechSupport  1 0.5869314      4788   3963.289 4009.289
```

Logistic Regression

Summary of our second logistic regression model

```
summary(step)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + PhoneService +
##     MultipleLines + InternetService + OnlineBackup + DeviceProtection +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges + Year, family = binomial(link = "logit
"),
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0395  -0.6602  -0.2795   0.6611   3.2193
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.838863   0.530232  -5.354 8.60e-08
## SeniorCitizen                    0.207458   0.100860   2.057  0.03970
## Dependents                      -0.255917   0.100533  -2.546  0.01091
## PhoneService                     0.827167   0.312842   2.644  0.00819
## MultipleLines                    0.598382   0.115028   5.202 1.97e-07
## InternetServiceFiber optic       2.566574   0.334424   7.675 1.66e-14
## InternetServiceNo               -2.655362   0.410543  -6.468 9.94e-11
## OnlineBackup                     0.178181   0.114907   1.551  0.12099
## DeviceProtection                 0.237463   0.116876   2.032  0.04218
## StreamingTV                      0.934428   0.163816   5.704 1.17e-08
## StreamingMovies                  0.931323   0.165648   5.622 1.88e-08
```

Logistic Regression

```
## PaymentMethodCredit card (automatic) -0.098532   0.137900  -0.715  0.47491
## PaymentMethodElectronic check         0.284582   0.114078   2.495  0.01261
## PaymentMethodMailed check             0.008648   0.139217   0.062  0.95047
## MonthlyCharges                       -1.982580   0.398856  -4.971 6.67e-07
## Year1-2                              -0.835135   0.116235  -7.185 6.73e-13
## Year2-3                              -1.270790   0.138407  -9.182  < 2e-16
## Year3-4                              -1.366834   0.154481  -8.848  < 2e-16
## Year4-5                              -1.592106   0.171416  -9.288  < 2e-16
## Year5-6                              -1.878456   0.203150  -9.247  < 2e-16
##
## (Intercept)                          ***
## SeniorCitizen                        *
## Dependents                           *
## PhoneService                         **
## MultipleLines                        ***
## InternetServiceFiber optic           ***
## InternetServiceNo                    ***
## OnlineBackup
## DeviceProtection                     *
## StreamingTV                          ***
## StreamingMovies                      ***
## ContractOne year                     ***
## ContractTwo year                     ***
```

Logistic Regression

```
## PaperlessBilling                          ***
## PaymentMethodCredit card (automatic)
## PaymentMethodElectronic check             *
## PaymentMethodMailed check
## MonthlyCharges                            ***
## Year1-2                                   ***
## Year2-3                                   ***
## Year3-4                                   ***
## Year4-5                                   ***
## Year5-6                                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5622.1  on 4810  degrees of freedom
## Residual deviance: 3963.3  on 4788  degrees of freedom
## AIC: 4009.3
##
## Number of Fisher Scoring iterations: 6
```

Logistic Regression

Evaluate the model

```
predict_train2 <- predict(step, newdata=test_data, type="response")
predict_train2 <- ifelse(predict_train2 > 0.5, 1, 0)
predict_error2 <- mean(predict_train2 != test_data$Churn)
model2 <- 1 - predict_error2
print(model2)
```

```
## [1] 0.7996398
```

Logistic Regression

Confusion matrix

```
table(ActualResult = test_data$Churn, Prediction = predict_train > 0.5)
```

```
##              Prediction
## ActualResult FALSE TRUE
##            0  1488  168
##            1   280  285
```

```
table(ActualResult = test_data$Churn, Prediction = predict_train2 > 0.5)
```

```
##              Prediction
## ActualResult FALSE TRUE
##            0  1489  167
##            1   278  287
```

Conclusion

Seyoung Jung
sjung.stat@gmail.com
https://github.com/sjung-stat/Telco
https://linkedin.com/in/sjung-stat