Week 11\_Helee

Helee Rana

5/4/2020

## Prediction about the customer will continue the Phone service with the company or not

require(tidyverse)

## Loading required package: tidyverse

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

## ✓ ggplot2 3.2.1 ✓ purrr 0.3.3  
## ✓ tibble 2.1.3 ✓ dplyr 0.8.4  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0

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

churn=read.csv("/Users/heleerana/Desktop/DATA MINING Prof. YEGIN/WA\_Fn-UseC\_-Telco-Customer-Churn.csv")  
head(churn)

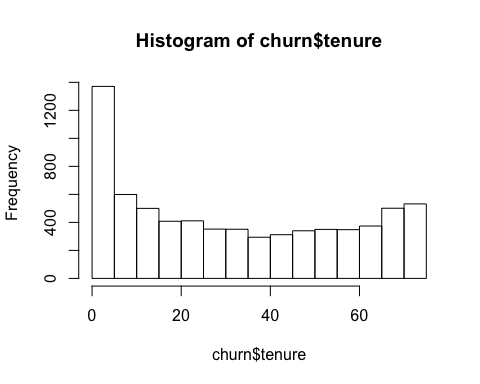
## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## 1 7590-VHVEG Female 0 Yes No 1 No  
## 2 5575-GNVDE Male 0 No No 34 Yes  
## 3 3668-QPYBK Male 0 No No 2 Yes  
## 4 7795-CFOCW Male 0 No No 45 No  
## 5 9237-HQITU Female 0 No No 2 Yes  
## 6 9305-CDSKC Female 0 No No 8 Yes  
## MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection  
## 1 No phone service DSL No Yes No  
## 2 No DSL Yes No Yes  
## 3 No DSL Yes Yes No  
## 4 No phone service DSL Yes No Yes  
## 5 No Fiber optic No No No  
## 6 Yes Fiber optic No No Yes  
## TechSupport StreamingTV StreamingMovies Contract PaperlessBilling  
## 1 No No No Month-to-month Yes  
## 2 No No No One year No  
## 3 No No No Month-to-month Yes  
## 4 Yes No No One year No  
## 5 No No No Month-to-month Yes  
## 6 No Yes Yes Month-to-month Yes  
## PaymentMethod MonthlyCharges TotalCharges Churn  
## 1 Electronic check 29.85 29.85 No  
## 2 Mailed check 56.95 1889.50 No  
## 3 Mailed check 53.85 108.15 Yes  
## 4 Bank transfer (automatic) 42.30 1840.75 No  
## 5 Electronic check 70.70 151.65 Yes  
## 6 Electronic check 99.65 820.50 Yes

## Since Senior citizen columnis in 0s and 1s (integer), we will need to convert to character

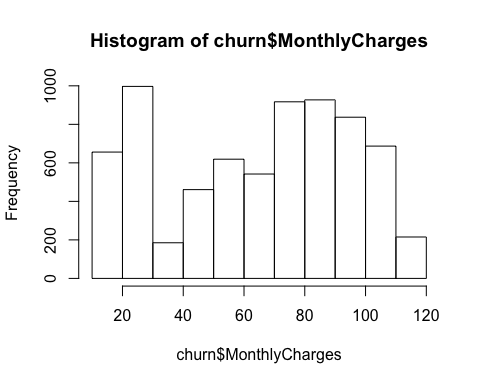
churn=churn%>%mutate(SeniorCitizen=as.character(SeniorCitizen))  
  
## We dont require customer id  
  
churn=churn%>%select(-customerID)

## Now we have tenure and other 3 columns as integer value, to implement naive bayes, the integer values in a column should be normally distributed. So lets check whether tenure other 3 columns are normally distributed or not

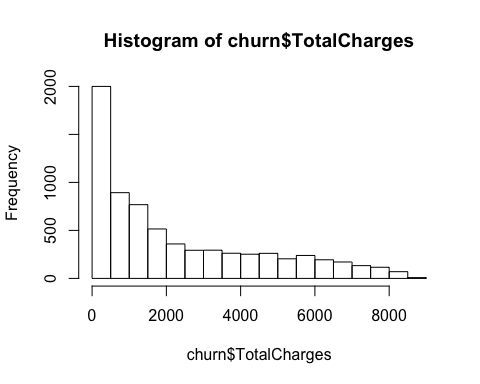
hist(churn$tenure) ## Candidate for binning



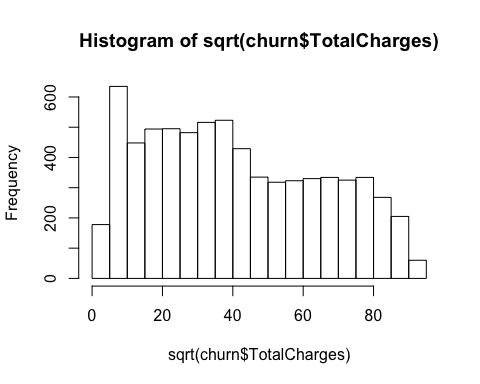
## We can see that the graph is not normally distributed  
## Now two options are to ignore the tenure column OR binning  
  
hist(churn$MonthlyCharges) ## Candidate for binning



## Same for MonthlyCharges  
  
hist(churn$TotalCharges)



## The graph seems exponential, Therefore now we take square root of the values  
hist(sqrt(churn$TotalCharges)) ## sqrt transformation



## Binning, that means dividing the values into two  
 ## one method   
##cbind(churn$tenure, as.factor(cut(churn$tenure, 2)))  
  
## second method if we want to insert labels as low and high  
##churn%>%mutate(tenure=cut(tenure,2, labels=c('low', 'high')))  
  
## Now we apply the second method to tenure, monthlycharges  
churn2=churn%>%mutate(tenure=cut(tenure,2), MonthlyCharges=cut(MonthlyCharges, 4), TotalCharges=sqrt(TotalCharges))

## Model Building and evaluation

set.seed(1234)  
n=nrow(churn2)  
training=runif(n)>.3  
training\_data=churn2[training,]  
testing\_data=churn2[!training,]  
  
library(e1071)  
require(caret)

## Loading required package: caret

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

nb\_model=naiveBayes(training\_data[-20], as.factor(training\_data$Churn))

## prediction

predictions=predict(nb\_model, testing\_data[-20], laplace=.1)

## Confusion Matrix

confusionMatrix(as.factor(predictions), as.factor(testing\_data$Churn))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1095 110  
## Yes 453 433  
##   
## Accuracy : 0.7308   
## 95% CI : (0.7112, 0.7497)  
## No Information Rate : 0.7403   
## P-Value [Acc > NIR] : 0.8467   
##   
## Kappa : 0.4189   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.7074   
## Specificity : 0.7974   
## Pos Pred Value : 0.9087   
## Neg Pred Value : 0.4887   
## Prevalence : 0.7403   
## Detection Rate : 0.5237   
## Detection Prevalence : 0.5763   
## Balanced Accuracy : 0.7524   
##   
## 'Positive' Class : No   
##