D208 Logistic Regression

Adam Eyerman

2022-07-28

# Part I: Research Question

**A1)** **Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.**

Can we predict the likelihood of churn using explanatory variables in the given data set? If so, which variables contribute the most to this prediction?

**A2)** **Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.**

The objective of the data analysis is to be able to predict the likelihood of churn for a customer using a variety of explanatory variables. If a logistic regression equation can predict the probability of churn within a reasonable certainty, then the stakeholders will benefit from understanding the variables behind why customers choose to terminate their accounts. Adjustments can then be made to any positive or negative correlations for the variables within the regression model to maximize the likelihood of customer retention and avoiding churn.

# Part II: Method Justification

**B1)** **Summarize the assumptions of a logistic regression model.**

According to (“The 6 Assumptions of Logistic Regression (With Examples)”,2020), a logistic regression model has 6 asssumptions:

* the response variable is binary
* the observations are independent
* no multicollinearity among explanatory variables
* no extreme outliers
* linear relationship between explanatory variables and logit of response variable
* sample size is sufficiently large

The response variable needs to have 2 outcomes to function within a logistic regression. In this case, Churn data is either Yes or No which is valid. The next two assumptions involve the relation of explanatory variables to each other. Each needs to be independent. Extreme outliers need to be removed to create valid logistic model. An established linear relationship must exist between explanatory and logit of response variable. This requires each explanatory variable to have a linear relationship with the probability (0-1) of the response variable. The sample size must be sufficient. The data set being used has 10,000 observations and will suffice.

**B2)** **Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.**

I will be using R and RStudio for my analysis using multiple logistic regression. According to (“R: What Is R?”, n.d.) R provides a wide variety of statistical and graphical techniques and is highly extensible. This includes linear and non-linear models and the ability to utilize extensions through packages to further its capabilities. R offers an open source and easy to navigate environment for data wrangling, manipulating, cleaning, and analyzing. My analysis will involve all these tasks and R serves as great tool to gather data, prepare it for analysis, and run predictive models. Also, the R-Markdown feature (RMD) is an efficient way to organize all aspects of the data analysis process and creates either an HTML, PDF, or Word document to present the analysis.

**B3) Explain why multiple regression is an appropriate technique to analyze the research question summarized in Part I.**

According to (Swaminathan, 2019), logistic regression is used when the dependent variable is categorical. The research question is attempting to determine the probability of Churn which is a categorical variable with response of “Yes” or “No”. This makes the dependent variable a binomial. Using logistic regression is ideal for determining the answer to the research question. By using binary logistic regression, the research question can be evaluated by forming a logistic regression model using a variety of explanatory variables to determine the probability of Churn.

# Part III: Data Preparation

**C1) Describe your data preparation goals and the data manipulations that will be used to achieve the goals.**

The main objective of the data preparation process is to create a clean and reliable data set of predictor variables and the target variable. The first step will be to upload the churn data set and evaluate the columns for any NULL or missing values. If found, then an imputation process will be done to normalize the data using central tendencies of mean, median or mode. Next would be to examine for outliers in the continuous variables. If found, this value would be removed. Also, the survey response categories (Item1-8) will be altered to their specific titles for more clarity. I will convert the binary and categorical variables into numeric values to utilize them for the multiple logistic regression equation. For example, the Churn variable consists of “Yes” and “No” for values. I would replace each value type with a number (0,1) to use the variable in the regression model. Once prepared, I will create a new data set (churn\_prepared) that will be used for developing the logistic regression model.

**C2)** **Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.**

The churn\_clean data set has 50 variables and 10,000 observations. For the research question, the data set will be reduced to 34 variables for the analysis. The variables of CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital, PaymentMethod were removed from the data set. The 34 remaining variables are numeric, integer, and categorical data. The categorical variables were converted to integer for the logistic regression model. These include: Gender, Churn, Techie, Contract, Port\_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling. The Gender variable has 3 distinct values and will be converted to 0, 1, 2 respectively while the remaining variables will be converted to 0 or 1 based on the Yes or No response value. The numeric variables will be Tenure, Outage\_sec\_perweek, MonthlyCharge, Bandwidth\_GB\_Year, and Income. The integer variables will be Children, Age, Yearly\_equip\_failure, Email, Contacts, Timely\_response, Timely\_fixes, Timely\_replacements, Reliability, Options, Respectful\_response, Courteous\_exchange, and Active\_listening.

Preparation: I will gather the summary statistics by running the summary() function, which will acquire the Minimum, 1st Quartile, Median, Mean, 3rd Quartile, and Maximum for each variable. Both univariate and bivariate visualizations will be analyzed. This will provide information about the distribution of the variables. The variables need a normal distribution to be effective in the logistic regression model.

Assessment: After assessing the summary statistics of the predictor variables there were no issues with distribution. Details shown in C3 and C4.

**C3)** **Explain the steps used to prepare the data for the analysis, including the annotated code.**

The data set was prepared using the following steps:

* Load the data set “churn\_clean.csv”
* Install package “tidyverse” and load library accordingly
* Check for NULL or missing values
* Rename columns (Item1-8) to respective names for more clarity
* Change categorical variables to integer
* Replace categorical variables with integer version
* Remove unnecessary variables that will not be included in linear regression model
* View summary statistics using
* Display univariate visualizations for distribution analysis
* Display bivariate visualization for distribution analysis with target variable

#Install package “tidyverse” install.packages(“tidyverse”) #Install packages needed for Confusion Matrix install.packages(“ISLR”) install.packages(“caret”) install.packages(“InformationValue”)

#Load libraries  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.6 v purrr 0.3.4  
## v tibble 3.1.7 v dplyr 1.0.9  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(InformationValue)  
library(ISLR)  
  
#Load the data set "churn\_clean.csv"  
churn\_clean <- read.csv('churn\_clean.csv')  
  
#Check for NULL or missing values  
colSums(is.na(churn\_clean))

## CaseOrder Customer\_id Interaction   
## 0 0 0   
## UID City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## TimeZone Job Children   
## 0 0 0   
## Age Income Marital   
## 0 0 0   
## Gender Churn Outage\_sec\_perweek   
## 0 0 0   
## Email Contacts Yearly\_equip\_failure   
## 0 0 0   
## Techie Contract Port\_modem   
## 0 0 0   
## Tablet InternetService Phone   
## 0 0 0   
## Multiple OnlineSecurity OnlineBackup   
## 0 0 0   
## DeviceProtection TechSupport StreamingTV   
## 0 0 0   
## StreamingMovies PaperlessBilling PaymentMethod   
## 0 0 0   
## Tenure MonthlyCharge Bandwidth\_GB\_Year   
## 0 0 0   
## Item1 Item2 Item3   
## 0 0 0   
## Item4 Item5 Item6   
## 0 0 0   
## Item7 Item8   
## 0 0

#Rename columns (Item1-8) to respective names for more clarity  
churn\_updated <- churn\_clean %>% rename(Timely\_response = Item1, Timely\_fixes = Item2, Timely\_replacements = Item3, Reliability = Item4, Options = Item5, Respectful\_response = Item6, Courteous\_exchange = Item7, Active\_listening = Item8)   
  
#Change categorical variables to integer  
churn\_new <- churn\_clean$Churn # Replicating vector  
churn\_new <- as.character(churn\_new) # Converting factor to character  
churn\_new[churn\_new == "Yes"] <- 1 # Replacing Yes by 1  
churn\_new[churn\_new == "No"] <- 0 # Replacing No by 0  
churn\_new <- as.integer(churn\_new) # Converting character to numeric  
  
gender\_new <- churn\_clean$Gender # Replicating vector  
gender\_new <- as.character(gender\_new) # Converting factor to character  
gender\_new[gender\_new == "Male"] <- 2 # Replacing Yes by 2  
gender\_new[gender\_new == "Female"] <- 1 # Replacing No by 1  
gender\_new[gender\_new == "Nonbinary"] <- 0 # Replacing No by 0  
gender\_new <- as.integer(gender\_new) # Converting character to numeric  
  
  
techie\_new <- churn\_clean$Techie # Replicating vector  
techie\_new <- as.character(techie\_new) # Converting factor to character  
techie\_new[techie\_new == "Yes"] <- 1 # Replacing Yes by 1  
techie\_new[techie\_new == "No"] <- 0 # Replacing No by 0  
techie\_new <- as.integer(techie\_new) # Converting character to numeric  
   
  
paperless\_billing\_new <- churn\_clean$PaperlessBilling # Replicating vector  
paperless\_billing\_new <- as.character(paperless\_billing\_new)# Converting factor to character  
paperless\_billing\_new[paperless\_billing\_new == "Yes"] <- 1 # Replacing Yes by 1  
paperless\_billing\_new[paperless\_billing\_new == "No"] <- 0 # Replacing No by 0  
paperless\_billing\_new <- as.integer(paperless\_billing\_new)# Converting character to numeric  
   
  
streaming\_movies\_new <- churn\_clean$StreamingMovies # Replicating vector  
streaming\_movies\_new <- as.character(streaming\_movies\_new) # Converting factor to character  
streaming\_movies\_new[streaming\_movies\_new == "Yes"] <- 1 # Replacing Yes by 1  
streaming\_movies\_new[streaming\_movies\_new == "No"] <- 0 # Replacing No by 0  
streaming\_movies\_new <- as.integer(streaming\_movies\_new) # Converting character to numeric  
  
  
streaming\_tv\_new <- churn\_clean$StreamingTV # Replicating vector  
streaming\_tv\_new <- as.character(streaming\_tv\_new) # Converting factor to character  
streaming\_tv\_new[streaming\_tv\_new == "Yes"] <- 1 # Replacing Yes by 1  
streaming\_tv\_new[streaming\_tv\_new == "No"] <- 0 # Replacing No by 0  
streaming\_tv\_new <- as.integer(streaming\_tv\_new) # Converting character to numeric  
   
  
techsupport\_new <- churn\_clean$TechSupport # Replicating vector  
techsupport\_new <- as.character(techsupport\_new) # Converting factor to character  
techsupport\_new[techsupport\_new == "Yes"] <- 1 # Replacing Yes by 1  
techsupport\_new[techsupport\_new == "No"] <- 0 # Replacing No by 0  
techsupport\_new <- as.integer(techsupport\_new) # Converting character to numeric  
  
  
device\_protection\_new <- churn\_clean$DeviceProtection # Replicating vector  
device\_protection\_new <- as.character(device\_protection\_new) # Converting factor to character  
device\_protection\_new[device\_protection\_new == "Yes"] <- 1# Replacing Yes by 1  
device\_protection\_new[device\_protection\_new == "No"] <- 0 # Replacing No by 0  
device\_protection\_new <- as.integer(device\_protection\_new) # Converting character to numeric  
   
  
online\_backup\_new <- churn\_clean$OnlineBackup # Replicating vector  
online\_backup\_new <- as.character(online\_backup\_new) # Converting factor to character  
online\_backup\_new[online\_backup\_new == "Yes"] <- 1 # Replacing Yes by 1  
online\_backup\_new[online\_backup\_new == "No"] <- 0 # Replacing No by 0  
online\_backup\_new <- as.integer(online\_backup\_new) # Converting character to numeric  
   
  
online\_security\_new <- churn\_clean$OnlineSecurity # Replicating vector  
online\_security\_new <- as.character(online\_security\_new) # Converting factor to character  
online\_security\_new[online\_security\_new == "Yes"] <- 1 # Replacing Yes by 1  
online\_security\_new[online\_security\_new == "No"] <- 0 # Replacing No by 0  
online\_security\_new <- as.integer(online\_security\_new) # Converting character to numeric  
   
  
multiple\_new <- churn\_clean$Multiple # Replicating vector  
multiple\_new <- as.character(multiple\_new) # Converting factor to character  
multiple\_new[multiple\_new == "Yes"] <- 1 # Replacing Yes by 1  
multiple\_new[multiple\_new == "No"] <- 0 # Replacing No by 0  
multiple\_new <- as.integer(multiple\_new) # Converting character to numeric  
  
  
phone\_new <- churn\_clean$Phone # Replicating vector  
phone\_new <- as.character(phone\_new) # Converting factor to character  
phone\_new[phone\_new == "Yes"] <- 1 # Replacing Yes by 1  
phone\_new[phone\_new == "No"] <- 0 # Replacing No by 0  
phone\_new <- as.integer(phone\_new) # Converting character to numeric  
  
  
tablet\_new <- churn\_clean$Tablet # Replicating vector  
tablet\_new <- as.character(tablet\_new) # Converting factor to character  
tablet\_new[tablet\_new == "Yes"] <- 1 # Replacing Yes by 1  
tablet\_new[tablet\_new == "No"] <- 0 # Replacing No by 0  
tablet\_new <- as.integer(tablet\_new) # Converting character to numeric  
  
  
port\_modem\_new <- churn\_clean$Port\_modem # Replicating vector  
port\_modem\_new <- as.character(port\_modem\_new) # Converting factor to character  
port\_modem\_new[port\_modem\_new == "Yes"] <- 1 # Replacing Yes by 1  
port\_modem\_new[port\_modem\_new == "No"] <- 0 # Replacing No by 0  
port\_modem\_new <- as.integer(port\_modem\_new) # Converting character to numeric  
   
  
contract\_new <- churn\_clean$Contract # Replicating vector  
contract\_new <- as.character(contract\_new) # Converting factor to character  
contract\_new[contract\_new == "Two Year"] <- 2 # Replacing Yes by 2  
contract\_new[contract\_new == "One year"] <- 1 # Replacing No by 1  
contract\_new[contract\_new == "Month-to-month"] <- 0 # Replacing No by 0  
contract\_new <- as.integer(contract\_new) # Converting character to numeric  
   
  
internet\_new <- churn\_clean$InternetService # Replicating vector  
internet\_new <- as.character(internet\_new) # Converting factor to character  
internet\_new[internet\_new == "Fiber Optic"] <- 2 # Replacing Yes by 2  
internet\_new[internet\_new == "DSL"] <- 1 # Replacing No by 1  
internet\_new[internet\_new == "None"] <- 0 # Replacing No by 0  
internet\_new <- as.integer(internet\_new) # Converting character to numeric  
   
  
#Replace categorical variables with integer version  
churn\_updated$Churn <- churn\_new  
churn\_updated$Contract <- contract\_new  
churn\_updated$DeviceProtection <- device\_protection\_new  
churn\_updated$Gender <- gender\_new  
churn\_updated$InternetService <- internet\_new  
churn\_updated$Multiple <- multiple\_new  
churn\_updated$OnlineSecurity <- online\_security\_new  
churn\_updated$OnlineBackup <- online\_backup\_new  
churn\_updated$PaperlessBilling <- paperless\_billing\_new  
churn\_updated$Phone <- phone\_new  
churn\_updated$Port\_modem <- port\_modem\_new  
churn\_updated$StreamingTV <- streaming\_tv\_new  
churn\_updated$StreamingMovies <- streaming\_movies\_new  
churn\_updated$Tablet <- tablet\_new  
churn\_updated$Techie <- techie\_new  
churn\_updated$TechSupport <- techsupport\_new  
  
#Remove unnecessary variables that will not be included in linear regression model  
churn\_prepared <- subset(churn\_updated, select = -c(CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital, PaymentMethod))  
   
#View summary statistics using  
summary(churn\_prepared)

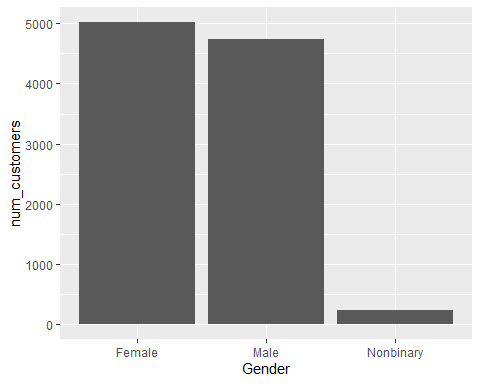
## Children Age Income Gender   
## Min. : 0.000 Min. :18.00 Min. : 348.7 Min. :0.000   
## 1st Qu.: 0.000 1st Qu.:35.00 1st Qu.: 19224.7 1st Qu.:1.000   
## Median : 1.000 Median :53.00 Median : 33170.6 Median :1.000   
## Mean : 2.088 Mean :53.08 Mean : 39806.9 Mean :1.451   
## 3rd Qu.: 3.000 3rd Qu.:71.00 3rd Qu.: 53246.2 3rd Qu.:2.000   
## Max. :10.000 Max. :89.00 Max. :258900.7 Max. :2.000   
## Churn Outage\_sec\_perweek Email Contacts   
## Min. :0.000 Min. : 0.09975 Min. : 1.00 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.: 8.01821 1st Qu.:10.00 1st Qu.:0.0000   
## Median :0.000 Median :10.01856 Median :12.00 Median :1.0000   
## Mean :0.265 Mean :10.00185 Mean :12.02 Mean :0.9942   
## 3rd Qu.:1.000 3rd Qu.:11.96949 3rd Qu.:14.00 3rd Qu.:2.0000   
## Max. :1.000 Max. :21.20723 Max. :23.00 Max. :7.0000   
## Yearly\_equip\_failure Techie Contract Port\_modem   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.398 Mean :0.1679 Mean :0.6986 Mean :0.4834   
## 3rd Qu.:1.000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :6.000 Max. :1.0000 Max. :2.0000 Max. :1.0000   
## Tablet InternetService Phone Multiple   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.0000 Median :1.000 Median :1.0000 Median :0.0000   
## Mean :0.2991 Mean :1.228 Mean :0.9067 Mean :0.4608   
## 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :2.000 Max. :1.0000 Max. :1.0000   
## OnlineSecurity OnlineBackup DeviceProtection TechSupport   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.000   
## Mean :0.3576 Mean :0.4506 Mean :0.4386 Mean :0.375   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.000   
## StreamingTV StreamingMovies PaperlessBilling Tenure   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. : 1.000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.: 7.918   
## Median :0.0000 Median :0.000 Median :1.0000 Median :35.431   
## Mean :0.4929 Mean :0.489 Mean :0.5882 Mean :34.526   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:61.480   
## Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :71.999   
## MonthlyCharge Bandwidth\_GB\_Year Timely\_response Timely\_fixes   
## Min. : 79.98 Min. : 155.5 Min. :1.000 Min. :1.000   
## 1st Qu.:139.98 1st Qu.:1236.5 1st Qu.:3.000 1st Qu.:3.000   
## Median :167.48 Median :3279.5 Median :3.000 Median :4.000   
## Mean :172.62 Mean :3392.3 Mean :3.491 Mean :3.505   
## 3rd Qu.:200.73 3rd Qu.:5586.1 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :290.16 Max. :7159.0 Max. :7.000 Max. :7.000   
## Timely\_replacements Reliability Options Respectful\_response  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :3.000 Median :3.000 Median :3.000 Median :3.000   
## Mean :3.487 Mean :3.498 Mean :3.493 Mean :3.497   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :8.000 Max. :7.000 Max. :7.000 Max. :8.000   
## Courteous\_exchange Active\_listening  
## Min. :1.00 Min. :1.000   
## 1st Qu.:3.00 1st Qu.:3.000   
## Median :4.00 Median :3.000   
## Mean :3.51 Mean :3.496   
## 3rd Qu.:4.00 3rd Qu.:4.000   
## Max. :7.00 Max. :8.000

**C4) Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.**

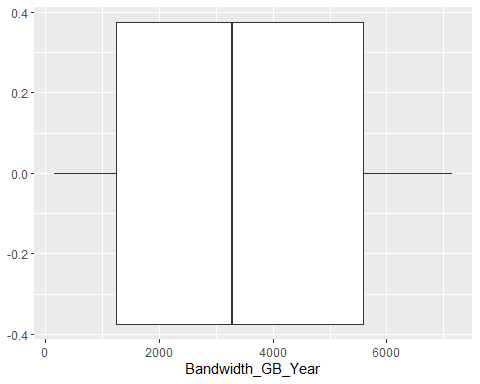
#Display univariate visualizations for distribution analysis  
num\_customers = 1  
  
#Churn  
churn\_plot <- ggplot(churn\_clean,  
 aes(Churn,num\_customers)) +  
 geom\_bar(stat = "identity")  
churn\_plot



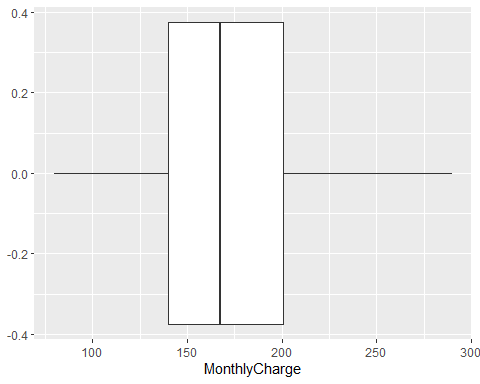
#Gender  
Gender\_plot <- ggplot(churn\_clean,  
 aes(Gender,num\_customers)) +  
 geom\_bar(stat = "identity")  
Gender\_plot



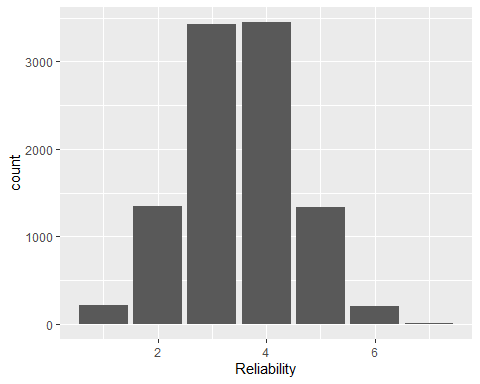
#Bandwidth  
Bandwidth\_plot <- ggplot(churn\_clean,  
 aes(Bandwidth\_GB\_Year)) +  
 geom\_boxplot()  
Bandwidth\_plot



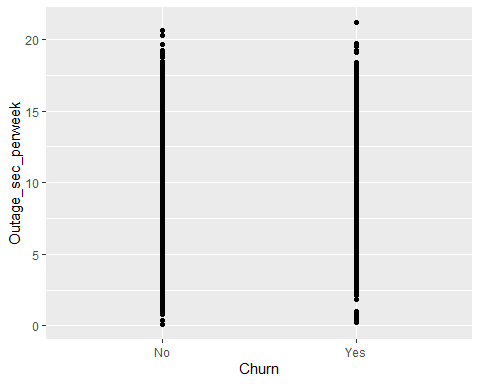
#MonthlyCharge  
MonthlyCharge\_plot <- ggplot(churn\_clean,  
 aes(MonthlyCharge)) +  
 geom\_boxplot()  
MonthlyCharge\_plot



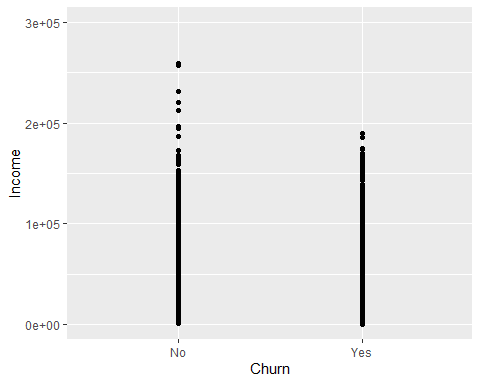
#Reliability  
Reliability\_plot <- ggplot(churn\_prepared,  
 aes(Reliability)) +  
 geom\_bar()  
Reliability\_plot



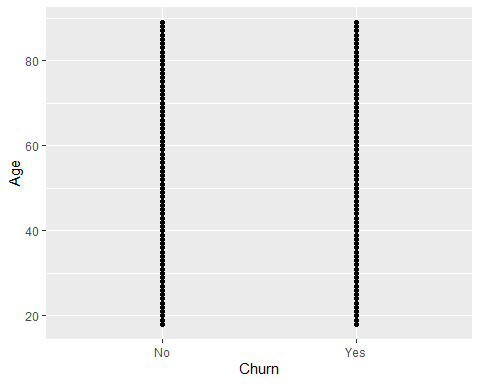
#Display bivariate visualization for distribution analysis with target variable  
  
#Outage(Bi)  
Outage\_biplot <- ggplot(churn\_clean,  
 aes(Churn, Outage\_sec\_perweek)) +  
 geom\_point()   
Outage\_biplot



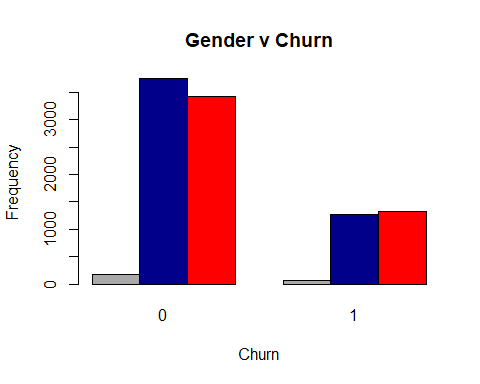
#Income(Bi)  
Income\_biplot <- ggplot(churn\_clean,  
 aes(Churn, Income)) +  
 geom\_point() +  
 ylim(0,300000)  
Income\_biplot



#Age(Bi)  
Age\_biplot <- ggplot(churn\_clean,  
 aes(Churn, Age)) +  
 geom\_point()   
Age\_biplot



#Table Creation for Biplot  
churn\_test <- churn\_new  
other\_table <- table(churn\_prepared$Gender, churn\_test)  
  
  
#Gender(Bi)  
Gender\_Churn <- barplot(other\_table,  
 main = "Gender v Churn",  
 xlab = "Churn", ylab = "Frequency",  
 col = c("darkgrey", "darkblue", "red"),  
 beside = TRUE) # Grouped bars



Gender\_Churn

## [,1] [,2]  
## [1,] 1.5 5.5  
## [2,] 2.5 6.5  
## [3,] 3.5 7.5

**C5) Provide a copy of the prepared data set.**

Prepared data set is attached as separate file in submission.

#Code of Prepared Data Set  
library(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

fwrite(churn\_prepared, "C:\\Users\\eyerm\\Dropbox\\PC\\Downloads\\churn\_prepared.csv")

# Part IV: Model Comparison and Analysis

**D1)** **Construct an initial logistic regression model from *all* predictors that were identified in Part C2.**

*churn\_new = -5.356 – 0.09\*Children + 0.126\*Age + 0.0000003\*Income + 0.005\*gender\_new -*

*0.39\*Tenure + 0.002\*Outage\_sec\_perweek – 0.006\*Email + 0.0045\*Contacts –*

*0.034\*Yearly\_equip\_failure + 1.003\*techie\_new + 0.154\*port\_modem\_new –*

*0.006\*tablet\_new – 0.489\* internet\_new -*

*0.31\*phone\_new + 0.286\*multiple\_new – 0.254\*online\_backup\_new –*

*0.444\*online\_security\_new +*

*0.92\*streaming\_movies\_new + 0.64\*streaming\_tv\_new + 0.0034\*Bandwidth\_GB\_Year +*

*0.032\* MonthlyCharge + 0.156\*paperless\_billing\_new – 1.938\*contract\_new –*

*0.316\*device\_protection\_new -*

*0.131\*techsupport\_new – 0.022\*Timely\_response + 0.015\*Timely\_fixes +*

*0.00432\* Timely\_replacements – 0.181\*Reliability – 0.028\*Options –*

*0.03\*Respectful\_response – 0.013\* Active\_listening + 0.0085\*Courteous\_exchange*

Initial Multiple Logistic Regression Model:

Logistic\_Model <- glm(churn\_new ~ Children + Age + Income + gender\_new + Tenure + Outage\_sec\_perweek + Email +  
 Contacts + Yearly\_equip\_failure + techie\_new + port\_modem\_new + tablet\_new +   
 internet\_new + phone\_new + multiple\_new + online\_backup\_new + online\_security\_new +   
 streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year + MonthlyCharge +   
 paperless\_billing\_new + contract\_new + device\_protection\_new + techsupport\_new +   
 Timely\_response + Timely\_fixes + Timely\_replacements + Reliability + Options +   
 Respectful\_response + Active\_listening + Courteous\_exchange, family = "binomial", data = churn\_prepared)  
   
  
summary(Logistic\_Model)

##   
## Call:  
## glm(formula = churn\_new ~ Children + Age + Income + gender\_new +   
## Tenure + Outage\_sec\_perweek + Email + Contacts + Yearly\_equip\_failure +   
## techie\_new + port\_modem\_new + tablet\_new + internet\_new +   
## phone\_new + multiple\_new + online\_backup\_new + online\_security\_new +   
## streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year +   
## MonthlyCharge + paperless\_billing\_new + contract\_new + device\_protection\_new +   
## techsupport\_new + Timely\_response + Timely\_fixes + Timely\_replacements +   
## Reliability + Options + Respectful\_response + Active\_listening +   
## Courteous\_exchange, family = "binomial", data = churn\_prepared)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9756 -0.2887 -0.0618 0.0875 3.4263   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.356e+00 5.460e-01 -9.811 < 2e-16 \*\*\*  
## Children -8.990e-02 1.843e-02 -4.879 1.07e-06 \*\*\*  
## Age 1.257e-02 1.958e-03 6.421 1.35e-10 \*\*\*  
## Income 2.923e-07 1.331e-06 0.220 0.826082   
## gender\_new 4.834e-02 6.983e-02 0.692 0.488784   
## Tenure -3.919e-01 1.774e-02 -22.092 < 2e-16 \*\*\*  
## Outage\_sec\_perweek 2.019e-03 1.258e-02 0.161 0.872464   
## Email -6.040e-03 1.234e-02 -0.490 0.624412   
## Contacts 4.464e-02 3.750e-02 1.190 0.233917   
## Yearly\_equip\_failure -3.360e-02 5.896e-02 -0.570 0.568718   
## techie\_new 1.003e+00 9.827e-02 10.207 < 2e-16 \*\*\*  
## port\_modem\_new 1.542e-01 7.479e-02 2.062 0.039229 \*   
## tablet\_new -6.003e-02 8.151e-02 -0.736 0.461487   
## internet\_new -4.894e-01 9.085e-02 -5.387 7.16e-08 \*\*\*  
## phone\_new -3.097e-01 1.281e-01 -2.419 0.015576 \*   
## multiple\_new 2.858e-01 1.686e-01 1.695 0.090120 .   
## online\_backup\_new -2.539e-01 1.289e-01 -1.970 0.048826 \*   
## online\_security\_new -4.435e-01 8.062e-02 -5.501 3.77e-08 \*\*\*  
## streaming\_movies\_new 9.156e-01 2.574e-01 3.558 0.000374 \*\*\*  
## streaming\_tv\_new 6.434e-01 2.176e-01 2.956 0.003116 \*\*   
## Bandwidth\_GB\_Year 3.437e-03 2.019e-04 17.022 < 2e-16 \*\*\*  
## MonthlyCharge 3.151e-02 4.572e-03 6.892 5.51e-12 \*\*\*  
## paperless\_billing\_new 1.557e-01 7.607e-02 2.047 0.040639 \*   
## contract\_new -1.938e+00 6.403e-02 -30.267 < 2e-16 \*\*\*  
## device\_protection\_new -3.162e-01 9.633e-02 -3.283 0.001027 \*\*   
## techsupport\_new -1.308e-01 9.664e-02 -1.354 0.175776   
## Timely\_response -2.224e-02 5.284e-02 -0.421 0.673798   
## Timely\_fixes 1.456e-02 5.016e-02 0.290 0.771566   
## Timely\_replacements 4.316e-03 4.556e-02 0.095 0.924538   
## Reliability -1.811e-02 4.050e-02 -0.447 0.654699   
## Options -2.797e-02 4.269e-02 -0.655 0.512258   
## Respectful\_response -3.042e-02 4.362e-02 -0.697 0.485587   
## Active\_listening -1.344e-02 3.909e-02 -0.344 0.730941   
## Courteous\_exchange 8.493e-03 4.159e-02 0.204 0.838205   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11564 on 9999 degrees of freedom  
## Residual deviance: 4615 on 9966 degrees of freedom  
## AIC: 4683  
##   
## Number of Fisher Scoring iterations: 7

pscl::pR2(Logistic\_Model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 0.6009286

**D2) Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.**

The research question is attempting to predict the probability of churn for any customer. The reduced model maintains the predictive power and retains the statistically significant variables leading to a more concise regression model with no loss of quality.

The initial model was reduced using the P-value significance metric of 0.05. All variables under that value were considered significant and kept in the reduced model. The remainder of the variables were removed. The model went from 33 variables to 16 variables. After the reduction, the model maintained its McFadden’s R-squared which was over 0.40 and confirmed the retained predictive power from the initial model.

**D3)** **Provide a reduced logistic regression model.**

*churn\_new = -5.982 – 0.936\*Children + 0.013\*Age - 0.398Tenure + 0.992\*techie\_new +*

*0.149\*port\_modem\_new -0.563\* internet\_new – 0.307\* phone\_new –*

*0.375\*online\_backup\_new – 0.465\*online\_security\_new + 0.631\*streaming\_movies\_new +*

*0.406\*streaming\_tv\_new + 0.0035\*Bandwidth\_GB\_Year +*

*0.036\*MonthlyCharge + 0.157\*paperless\_billing\_new - 1.94\*contract\_new -*

*0.3765\*device\_protection\_new*

Reduced Logistic Regression Model:

Reduced\_Logistic\_Model <- glm(churn\_new ~ Children + Age + Tenure + techie\_new + port\_modem\_new +   
 internet\_new + phone\_new + online\_backup\_new + online\_security\_new +   
 streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year + MonthlyCharge +   
 paperless\_billing\_new + contract\_new + device\_protection\_new, family = "binomial", data = churn\_prepared)  
  
  
summary(Reduced\_Logistic\_Model)

##   
## Call:  
## glm(formula = churn\_new ~ Children + Age + Tenure + techie\_new +   
## port\_modem\_new + internet\_new + phone\_new + online\_backup\_new +   
## online\_security\_new + streaming\_movies\_new + streaming\_tv\_new +   
## Bandwidth\_GB\_Year + MonthlyCharge + paperless\_billing\_new +   
## contract\_new + device\_protection\_new, family = "binomial",   
## data = churn\_prepared)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9597 -0.2892 -0.0612 0.0880 3.4212   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.9815301 0.2884953 -20.734 < 2e-16 \*\*\*  
## Children -0.0936626 0.0182648 -5.128 2.93e-07 \*\*\*  
## Age 0.0128083 0.0019459 6.582 4.64e-11 \*\*\*  
## Tenure -0.3984486 0.0173747 -22.933 < 2e-16 \*\*\*  
## techie\_new 0.9916452 0.0978798 10.131 < 2e-16 \*\*\*  
## port\_modem\_new 0.1493646 0.0745048 2.005 0.044988 \*   
## internet\_new -0.5628326 0.0615890 -9.139 < 2e-16 \*\*\*  
## phone\_new -0.3070495 0.1270427 -2.417 0.015653 \*   
## online\_backup\_new -0.3750053 0.0859388 -4.364 1.28e-05 \*\*\*  
## online\_security\_new -0.4653944 0.0793533 -5.865 4.50e-09 \*\*\*  
## streaming\_movies\_new 0.6315840 0.1250259 5.052 4.38e-07 \*\*\*  
## streaming\_tv\_new 0.4061122 0.1136615 3.573 0.000353 \*\*\*  
## Bandwidth\_GB\_Year 0.0035134 0.0001978 17.764 < 2e-16 \*\*\*  
## MonthlyCharge 0.0364434 0.0020518 17.762 < 2e-16 \*\*\*  
## paperless\_billing\_new 0.1573363 0.0757674 2.077 0.037841 \*   
## contract\_new -1.9376541 0.0640468 -30.254 < 2e-16 \*\*\*  
## device\_protection\_new -0.3765852 0.0791496 -4.758 1.96e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11564.4 on 9999 degrees of freedom  
## Residual deviance: 4630.5 on 9983 degrees of freedom  
## AIC: 4664.5  
##   
## Number of Fisher Scoring iterations: 7

pscl::pR2(Reduced\_Logistic\_Model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 0.5995893

**E1)** **Explain your data analysis process by comparing the initial and reduced logistic regression models, including the following elements:**

**•  the logic of the variable selection technique**

**•  the model evaluation metric**

The initial model was reduced using the P-value significance metric of 0.05. All variables under that value were considered significant and kept in the reduced model. The remainder of the variables were removed. The model went from 33 variables to 16 variables. After the reduction, the model maintained its McFadden’s R-squared which was over 0.40 and confirmed the retained predictive power from the initial model.

#McFadden Values  
pscl::pR2(Logistic\_Model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 0.6009286

pscl::pR2(Reduced\_Logistic\_Model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 0.5995893

#Summary of Models  
summary(Logistic\_Model)

##   
## Call:  
## glm(formula = Churn ~ Children + Age + Income + gender\_new +   
## Tenure + Outage\_sec\_perweek + Email + Contacts + Yearly\_equip\_failure +   
## techie\_new + port\_modem\_new + tablet\_new + internet\_new +   
## phone\_new + multiple\_new + online\_backup\_new + online\_security\_new +   
## streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year +   
## MonthlyCharge + paperless\_billing\_new + contract\_new + device\_protection\_new +   
## techsupport\_new + Timely\_response + Timely\_fixes + Timely\_replacements +   
## Reliability + Options + Respectful\_response + Active\_listening +   
## Courteous\_exchange, family = "binomial", data = churn\_prepared)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9756 -0.2887 -0.0618 0.0875 3.4263   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.356e+00 5.460e-01 -9.811 < 2e-16 \*\*\*  
## Children -8.990e-02 1.843e-02 -4.879 1.07e-06 \*\*\*  
## Age 1.257e-02 1.958e-03 6.421 1.35e-10 \*\*\*  
## Income 2.923e-07 1.331e-06 0.220 0.826082   
## gender\_new 4.834e-02 6.983e-02 0.692 0.488784   
## Tenure -3.919e-01 1.774e-02 -22.092 < 2e-16 \*\*\*  
## Outage\_sec\_perweek 2.019e-03 1.258e-02 0.161 0.872464   
## Email -6.040e-03 1.234e-02 -0.490 0.624412   
## Contacts 4.464e-02 3.750e-02 1.190 0.233917   
## Yearly\_equip\_failure -3.360e-02 5.896e-02 -0.570 0.568718   
## techie\_new 1.003e+00 9.827e-02 10.207 < 2e-16 \*\*\*  
## port\_modem\_new 1.542e-01 7.479e-02 2.062 0.039229 \*   
## tablet\_new -6.003e-02 8.151e-02 -0.736 0.461487   
## internet\_new -4.894e-01 9.085e-02 -5.387 7.16e-08 \*\*\*  
## phone\_new -3.097e-01 1.281e-01 -2.419 0.015576 \*   
## multiple\_new 2.858e-01 1.686e-01 1.695 0.090120 .   
## online\_backup\_new -2.539e-01 1.289e-01 -1.970 0.048826 \*   
## online\_security\_new -4.435e-01 8.062e-02 -5.501 3.77e-08 \*\*\*  
## streaming\_movies\_new 9.156e-01 2.574e-01 3.558 0.000374 \*\*\*  
## streaming\_tv\_new 6.434e-01 2.176e-01 2.956 0.003116 \*\*   
## Bandwidth\_GB\_Year 3.437e-03 2.019e-04 17.022 < 2e-16 \*\*\*  
## MonthlyCharge 3.151e-02 4.572e-03 6.892 5.51e-12 \*\*\*  
## paperless\_billing\_new 1.557e-01 7.607e-02 2.047 0.040639 \*   
## contract\_new -1.938e+00 6.403e-02 -30.267 < 2e-16 \*\*\*  
## device\_protection\_new -3.162e-01 9.633e-02 -3.283 0.001027 \*\*   
## techsupport\_new -1.308e-01 9.664e-02 -1.354 0.175776   
## Timely\_response -2.224e-02 5.284e-02 -0.421 0.673798   
## Timely\_fixes 1.456e-02 5.016e-02 0.290 0.771566   
## Timely\_replacements 4.316e-03 4.556e-02 0.095 0.924538   
## Reliability -1.811e-02 4.050e-02 -0.447 0.654699   
## Options -2.797e-02 4.269e-02 -0.655 0.512258   
## Respectful\_response -3.042e-02 4.362e-02 -0.697 0.485587   
## Active\_listening -1.344e-02 3.909e-02 -0.344 0.730941   
## Courteous\_exchange 8.493e-03 4.159e-02 0.204 0.838205   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11564 on 9999 degrees of freedom  
## Residual deviance: 4615 on 9966 degrees of freedom  
## AIC: 4683  
##   
## Number of Fisher Scoring iterations: 7

summary(Reduced\_Logistic\_Model)

##   
## Call:  
## glm(formula = churn\_new ~ Children + Age + Tenure + techie\_new +   
## port\_modem\_new + internet\_new + phone\_new + online\_backup\_new +   
## online\_security\_new + streaming\_movies\_new + streaming\_tv\_new +   
## Bandwidth\_GB\_Year + MonthlyCharge + paperless\_billing\_new +   
## contract\_new + device\_protection\_new, family = "binomial",   
## data = churn\_prepared)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9597 -0.2892 -0.0612 0.0880 3.4212   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.9815301 0.2884953 -20.734 < 2e-16 \*\*\*  
## Children -0.0936626 0.0182648 -5.128 2.93e-07 \*\*\*  
## Age 0.0128083 0.0019459 6.582 4.64e-11 \*\*\*  
## Tenure -0.3984486 0.0173747 -22.933 < 2e-16 \*\*\*  
## techie\_new 0.9916452 0.0978798 10.131 < 2e-16 \*\*\*  
## port\_modem\_new 0.1493646 0.0745048 2.005 0.044988 \*   
## internet\_new -0.5628326 0.0615890 -9.139 < 2e-16 \*\*\*  
## phone\_new -0.3070495 0.1270427 -2.417 0.015653 \*   
## online\_backup\_new -0.3750053 0.0859388 -4.364 1.28e-05 \*\*\*  
## online\_security\_new -0.4653944 0.0793533 -5.865 4.50e-09 \*\*\*  
## streaming\_movies\_new 0.6315840 0.1250259 5.052 4.38e-07 \*\*\*  
## streaming\_tv\_new 0.4061122 0.1136615 3.573 0.000353 \*\*\*  
## Bandwidth\_GB\_Year 0.0035134 0.0001978 17.764 < 2e-16 \*\*\*  
## MonthlyCharge 0.0364434 0.0020518 17.762 < 2e-16 \*\*\*  
## paperless\_billing\_new 0.1573363 0.0757674 2.077 0.037841 \*   
## contract\_new -1.9376541 0.0640468 -30.254 < 2e-16 \*\*\*  
## device\_protection\_new -0.3765852 0.0791496 -4.758 1.96e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11564.4 on 9999 degrees of freedom  
## Residual deviance: 4630.5 on 9983 degrees of freedom  
## AIC: 4664.5  
##   
## Number of Fisher Scoring iterations: 7

**E2) Provide the output and *any* calculations of the analysis you performed, including a confusion matrix.**

The confusion matrices for both the Initial and Reduced model displayed a similar output.

The predictions for the refined model were as follows:

* True Negative = 6901
* True Positive = 2805
* False Negative = 449
* False Positive = 565

The total misclassification rate was 0.265.

The predictions for the initial model were as follows:

* True Negative = 6901
* True Positive = 2802
* False Negative = 449
* False Positive = 568

The total misclassification rate was 0.265.

The percentage of individuals the model correctly predicted would churn is known as the sensitivity (“How to Create a Confusion Matrix in R (Step-by-Step)”, 2021). Both models had the same output.

The percentage of individuals the model correctly predicted would *not* churn at the specificity (“How to Create a Confusion Matrix in R (Step-by-Step)”, 2021). Both models had the same output.

#Initial Model Confusion Matrix  
predicted <- predict(Logistic\_Model, churn\_prepared, type="response")  
  
confusionMatrix(churn\_new, predicted)

## 0 1  
## 0 6901 568  
## 1 449 2082

sensitivity(churn\_new, predicted)

## [1] 0.7856604

specificity(churn\_new, predicted)

## [1] 0.9389116

misClassError(churn\_new, predicted, threshold="optimal")

## [1] 0.265

#Reduced Model Confusion Matrix  
predicted\_2 <- predict(Reduced\_Logistic\_Model, churn\_prepared, type="response")  
  
confusionMatrix(churn\_new, predicted\_2)

## 0 1  
## 0 6901 565  
## 1 449 2085

sensitivity(churn\_new, predicted\_2)

## [1] 0.7867925

specificity(churn\_new, predicted\_2)

## [1] 0.9389116

misClassError(churn\_new, predicted\_2, threshold="optimal")

## [1] 0.265

**E3)** **Provide the code used to support the implementation of the logistic regression models.**

All code shown above in D1-E2.

# Part V: Data Summary and Implications

**F1) Discuss the results of your data analysis, including the following elements:**

**•  a regression equation for the reduced model**

**•  an interpretation of coefficients of the statistically significant variables of the model**

**•  the statistical and practical significance of the model**

**•  the limitations of the data analysis**

***Reduced Regression Equation:***

*churn\_new = -5.982 – 0.936\*Children + 0.013\*Age - 0.398Tenure + 0.992\*techie\_new +*

*0.149\*port\_modem\_new -0.563\* internet\_new – 0.307\* phone\_new –*

*0.375\*online\_backup\_new – 0.465\*online\_security\_new + 0.631\*streaming\_movies\_new +*

*0.406\*streaming\_tv\_new + 0.0035\*Bandwidth\_GB\_Year +*

*0.036\*MonthlyCharge + 0.157\*paperless\_billing\_new - 1.94\*contract\_new -*

*0.3765\*device\_protection\_new*

The coefficients are showing their relationship to the target variable probability with either negative or positive correlation. For example, the variable of Age is positive and suggests that as age increases the probability of churn also rises. However, a variable such as Tenure has a negative coefficient which suggest the longer a customer stays with the company the less likely for churn. Knowing this relationship between predictor variables and the target variable is essential to determining an action step moving forward.

Instead of utilizing the R-squared value used in linear regression, the McFadden R-squared was produced to justify the significance or predictive power of the logistic model (“How to Perform Logistic Regression in R (Step-by-Step), 2021).

pscl::pR2(Reduced\_Logistic\_Model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 0.5995893

This value is over 0.40 which suggests a high predictive power and justifies the significance of the model.

This analysis is limited by the amount of data collected and the ability to update the data set over time. With a larger set of data and a recurring update the analysis may be more accurate and relevant.

**F2) Recommend a course of action based on your results.**

My recommended course of action would be to group the variables into positive or negative contributors to churn likelihood. This would establish a strategy to address the variables according to their relationship with the target variable and assist in the process to lower the churn rate. Understanding your customer and factors that lead to them terminating a contract is essential to maximizing retention and growth for the company.

Bandwidth and Monthly Charge may lead to higher churn as they rise. Perhaps lowering costs for high bandwidth output customers could lead to a churn reduction. Logically, the increase in Bandwidth would naturally lead to higher charge if prices were correlated to the amount of data being used. A deal for high usage customers may be necessary to negate this relationship to churn.

My last recommendation would be to continue to promote a variety of services as they seem to relate to lowering churn. Marketing a variety of services and focusing on customers adding services over time may lead to lower churn.

# Part VI: Demonstration

**G) Provide a Panopto video recording that includes *all* of the following elements:**

**•  a demonstration of the functionality of the code used for the analysis**

**•  an identification of the version of the programming environment**

**•  a comparison of the two logistic regression models you used in your analysis**

**•  an interpretation of the coefficients.**

[**Thu Jul 28 2022 2:04:30 PM (panopto.com)**](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a1d7690f-71fb-4805-9af7-aee0012a4cfd)

**H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.**

Z. (2020, October 13). *The 6 Assumptions of Logistic Regression (With Examples)*. Statology. https://www.statology.org/assumptions-of-logistic-regression/

*R: What is R?* (n.d.). R-Project. <https://www.r-project.org/about.html>

Swaminathan, S. (2019, January 18). *Logistic Regression — Detailed Overview - Towards Data Science*. Medium. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Z. (2021a, April 1). *How to Create a Confusion Matrix in R (Step-by-Step)*. Statology. <https://www.statology.org/confusion-matrix-in-r/>

Z. (2021, September 29). *How to Perform Logistic Regression in R (Step-by-Step)*. Statology. https://www.statology.org/logistic-regression-in-r/

**I) Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**

Cited throughout using APA format.