Analysis of Categorical Data - Assignment - Phase 2

MATH 1298 Analysis of Categorical Data Project Phase II

Amal Joy (s3644794) & Nupura Sanjay Sawle (s3639703)

14 October 2018

Contents

1 Introduction

The aim of this project is to build a customer churn model for a telecommunication company to predict the customers who are about to get churned so that they can implement different business strategies to retain those customers before they actually get churned. The tool that I am using for this analysis is R-studio. This project was be conducted in 2 different phase where I went through exploring different data analysis techniques to accurately predict the churned customers. Phase 1 of this project will included the detailed descriptive statistical analysis of the data by making use of various R packages to build relevant charts, graphs, and interactions etc. Data preprocessing was be done to clean and transform the data to suit the prediction model.

In this second phase of this project we are building the model after checking appropriate statistical procedures, the test of independence etc. The relevence and independence of different variables are explored using confidence intervals and hypothesis analysis. The dataset was completely analysed for different assumptions made for the purpose of this project.

2 Dataset source and description

The following packages are used in this report for data preparation and data modeling.

```
library(dplyr)
library(knitr)
library(kableExtra)
library(ggplot2)
library(package = binom)
library(caret)
library(MASS)
library(car)
```

The data was read into a data file named 'telcom_churn'. Null values are replaced with 'NA' while reading the file.

```
setwd("/Users/amaljoy/Study/Categorcal Data/Assignment 2/")
telcom_churn <- read.csv(
  "/Users/amaljoy/Study/Categorcal Data/Assignment 2/Telco-Customer-Churn.csv",
  header=T,na.strings=c("","NA")) # Reading the data

dim(telcom_churn) # dimensions of the dataset
```

```
## [1] 7043   21
```

The dataset consists of 21 variables and 7043 observations. Each row in the dataset is the attributes associated to a customer, each column contains customer's attributes. The customer attributes are provided below: * customerID (Unique customer identification)
* gender (female, male)
* SeniorCitizen (Whether the customer is a senior citizen or not (1, 0))
* Partner (Whether the customer has a partner or not (Yes, No))
* Dependents (Whether the customer has dependents or not (Yes, No))
* tenure (Number of months the customer has stayed with the company)
* PhoneService (Whether the customer has a phone service or not (Yes, No))
* MultipleLines (Whether the customer has multiple lines r not (Yes, No, No phone service)
* InternetService (Customer's internet service provider (DSL, Fiber optic, No)
* OnlineSecurity (Whether the customer has online security or not (Yes, No, No internet service)
* OnlineBackup (Whether the customer has an online backup or not (Yes, No, No internet service)

* DeviceProtection (Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport (Whether the customer has tech support or not (Yes, No, No internet service)
* streamingTV (Whether the customer has streaming TV or not (Yes, No, No internet service)
* streamingMovies (Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract (The contract term of the customer (Month-to-month, One year, Two year)
* Paperless billing (Whether the customer has paperless billing or not (Yes, No))
* PaymentMethod (The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)))
* MonthlyCharges (The amount charged to the customer monthly - numeric)
* TotalCharges (The total amount charged to the customer - numeric)
* Churn (Whether the customer churned or not (Yes or No))

The attribute churn will be the target variable. Given below is the first 5 observations in the dataset.

Table 1: Head of the data

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
7590-VHVEG	Female	0	Yes	No	1	No	No phone service
5575-GNVDE	Male	0	No	No	34	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No
7795-CFOCW	Male	0	No	No	45	No	No phone service
9237-HQITU	Female	0	No	No	2	Yes	No
9305-CDSKC	Female	0	No	No	8	Yes	Yes

InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV
DSL	No	Yes	No	No	No
DSL	Yes	No	Yes	No	No
DSL	Yes	Yes	No	No	No
DSL	Yes	No	Yes	Yes	No
Fiber optic	No	No	No	No	No
Fiber optic	No	No	Yes	No	Yes

StreamingMovi	Contract	PaperlessBilling	PaymentMetho	MonthlyCharge	TotalCharges	Churn
No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	One year	No	Mailed check	56.95	1889.50	No
No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
Yes	Month-to-month	Yes	Electronic check	99.65	820.50	Yes

2.1 Summary Statistics

Summary of the dataset is given below. It shows the summary of each variable including the levels in case of factors; range and central tendency values in case of numerical values.

```
# Summary of the data
summary(telcom_churn)
```

```
##      customerID      gender     SeniorCitizen    Partner    Dependents
##  0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933
##  0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110
##  0004-TLHLJ:   1                 Median :0.0000
##  0011-IGKFF:   1                 Mean   :0.1621
##  0013-EXCHZ:   1                 3rd Qu.:0.0000
##  0013-MHZWF:   1                 Max.   :1.0000
##  (Other)   :7037
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##
##             OnlineSecurity              OnlineBackup
##  No                 :3498   No                 :3088
##  No internet service:1526   No internet service:1526
##  Yes                :2019   Yes                :2429
##
##
##
##
##            DeviceProtection              TechSupport
##  No                 :3095    No                 :3473
##  No internet service:1526    No internet service:1526
##  Yes                :2422    Yes                :2044
##
##
##
##
##              StreamingTV              StreamingMovies
##  No                 :2810   No                 :2785
##  No internet service:1526   No internet service:1526
##  Yes                :2707   Yes                :2732
##
##
##
##
##            Contract    PaperlessBilling                   PaymentMethod
##  Month-to-month:3875   No :2872         Bank transfer (automatic):1544
##  One year      :1473   Yes:4171         Credit card (automatic)  :1522
##  Two year      :1695                    Electronic check         :2365
##                                         Mailed check             :1612
##
##
##
##  MonthlyCharges    TotalCharges    Churn
##  Min.   : 18.25   Min.   :  18.8   No :5174
##  1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
##  Median : 70.35   Median :1397.5
##  Mean   : 64.76   Mean   :2283.3
##  3rd Qu.: 89.85   3rd Qu.:3794.7
```

```
## Max.   :118.75   Max.   :8684.8
##                  NA's   :11
```

3 Data Preprocessing

All of the data preprocessing tasks completed in the phase I of this project is briefed here. ‘SeniorCitizen’ has to be converted to a factor. It is labeled as ‘Yes’, and ‘No’.

```
telcom_churn$SeniorCitizen= factor(telcom_churn$SeniorCitizen, c(0,1), labels = c('No','Yes'),
                                   ordered = is.ordered(telcom_churn))
```

We cannot consider the attributes of a recent customer to train the model. For the stability of the model and better accuracy, we are considering only those customers who are customers of telco for at least 1 year, i.e 12 months.

```
telcom_churn <- subset(telcom_churn,telcom_churn$tenure>12) # creating new subset
```

Now we have 4857 churned/active customers who are with telco for more than one year.

To reduce the Curse of Dimensionality and the time required to train the algorithm, we chose to factorise the variable ‘tenure’ in to 5 different groups; 13-24 months, 25-36 months, 37-48 months, 49-60 months, and 61-72 months.

```
telcom_churn$tenure <- cut(telcom_churn$tenure,breaks=5,dig.lab=2,labels=2:6)
```

4 Modeling

4.1 Confident Intervals

Before performing the regression, lets analyse the target variable. The probability of any customer getting churned is to be calculated and its confidence limit is to be determined.

```
# as.numeric(table(telcom_churn$Churn)[2])
w<-sum(telcom_churn$Churn=="Yes") # Number of customers who got churned
n<-length(telcom_churn$Churn) # Total number of customers
alpha<-0.05 # 95% Confidence
pi.hat<-w/n
pi.hat # Point probability of getting churned
```

```
## [1] 0.1712992
```

```
binom.confint(x = w, n = n, conf.level = 1-alpha, methods = "all") # Confident interval methods
```

```
##           method   x    n      mean     lower     upper
## 1  agresti-coull 832 4857 0.1712992 0.1609608 0.1821571
## 2     asymptotic 832 4857 0.1712992 0.1607032 0.1818951
## 3          bayes 832 4857 0.1713668 0.1608118 0.1819977
## 4        cloglog 832 4857 0.1712992 0.1608477 0.1820345
## 5          exact 832 4857 0.1712992 0.1608002 0.1821948
## 6          logit 832 4857 0.1712992 0.1609616 0.1821565
## 7         probit 832 4857 0.1712992 0.1609129 0.1821043
## 8        profile 832 4857 0.1712992 0.1608779 0.1820671
## 9            lrt 832 4857 0.1712992 0.1608786 0.1820826
## 10     prop.test 832 4857 0.1712992 0.1608636 0.1822593
## 11        wilson 832 4857 0.1712992 0.1609640 0.1821539
```

With 95% of confidence, the true probability of a customer getting churned, given by various methods are between the lower and upper limits provided in the above table. It can be noted that almost all of the methods give nearly the same amount of confidence limit. However, since the number of observations are well beyond 40, we can go for Agresti-Coull confidence limits for which 95% confidence interval is $0.161 < \pi < 0.182$.

4.2 Contigency table

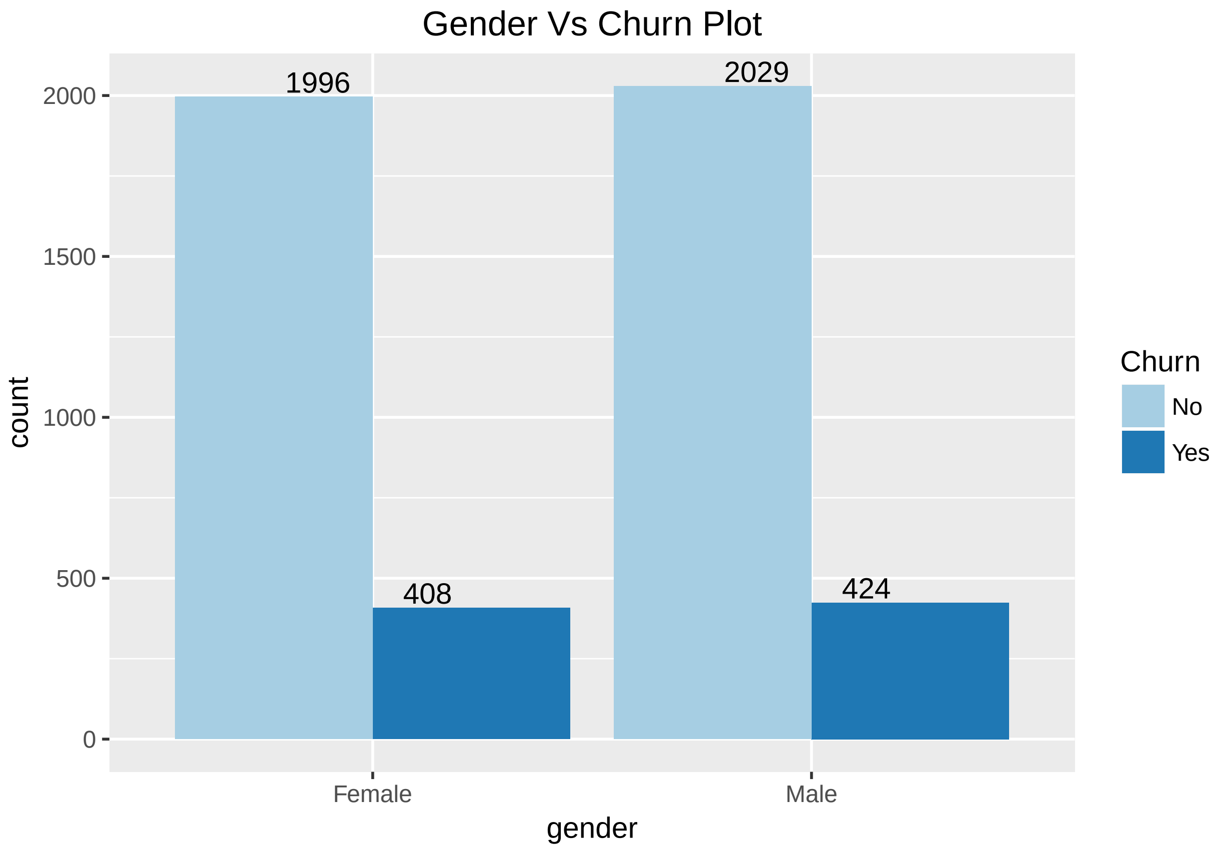
All the variables that we have in the dataset may not be relevent in deciding the churn tendency of the customer. Some of them may hardly give any information to the model. To reduce the curse of dimentionality, we may remove the variables that are not relevant from the analysis. For this purpose we are looking for the confidence limit of difference between two probabilities.

4.2.1 Gender

We have noticed from the gender vs churn bar plot from the phase I that plot is identical for male and female customers. The plot is given below.

```
# Analysing Gender variable
ggplot(telcom_churn, aes(gender, ..count..,fill = Churn)) +
geom_bar(stat="count", position = "dodge") +
scale_fill_brewer(palette="Paired") +
ggtitle("Gender Vs Churn Plot") +
```

```
theme(plot.title = element_text(hjust = 0.5)) +
geom_text(aes(label=..count..),
          stat="count",position=position_dodge(0.5),vjust=-0.2)
```



If there is not much difference between the male and female customers in getting churned this variable can hardly provide any information about the churn tendency of the customers. A contigency table is constructed for the variables gender and churn.

```
YF <- sum(telcom_churn$Churn=="Yes" & telcom_churn$gender=="Female")
YM <- sum(telcom_churn$Churn=="Yes" & telcom_churn$gender=="Male")
NF <- sum(telcom_churn$Churn=="No" & telcom_churn$gender=="Female")
NM <- sum(telcom_churn$Churn=="No" & telcom_churn$gender=="Male")

c.table<-array(data = c(YF,YM,NF,NM),
               dim = c(2,2),
               dimnames = list(gender = c("Female", "Male"),
               Churn = c("Yes", "No")))
c.table
```

```
##         Churn
## gender   Yes   No
##   Female 408 1996
##   Male   424 2029
```

```
ifelse(sum(c.table)==length(telcom_churn$Churn),
       "Contingency table includes all the observations",
       "Contigency Table Made is not correct")
```

```
## [1] "Contingency table includes all the observations"
```

```r
# Find the estimated pi^j
pi.hat.table<-c.table/rowSums(c.table)
pi.hat.table
```

```
##         Churn
## gender         Yes        No
##   Female 0.1697171 0.8302829
##   Male   0.1728496 0.8271504
```

16.97 % of female customers got churned and 17.28 % male customers got churned. While 83.03 % females are still active customers, 82.72 % of the males customers are active with Telco.

So now we have to check if there is a difference between the proportion of males and females who got churned. For that purpose we will consider both Wald and Agresti-Caffo confidence intervals.

```r
# Confidence interval for difference of two probabilities
alpha<-0.05
pi.hat1<-pi.hat.table[1,1] # Proportion of female customers who got churned
pi.hat2<-pi.hat.table[2,1] # Proportion of male customers who got churned

# Wald CI
var.wald<-pi.hat1*(1-pi.hat1) / sum(c.table[1,]) + pi.hat2*(1-pi.hat2) / sum(c.table[2,])
wal1 <- pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald)
wal1
```

```
## [1] -0.02432371  0.01805884
```

Therefore 95% Wald confidence interval is -0.0243 $< \pi_1 - \pi_2 <$ 0.0181. Since this interval include zero, there is no sufficient evidence to indicate the difference in the proportions of male and female customers who got churned.

```r
# Agresti-Caffo CI
pi.tilde1<-(c.table[1,1]+1)/(sum(c.table[1,])+2)
pi.tilde2<-(c.table[2,1]+1)/(sum(c.table[2,])+2)
var.AC<-pi.tilde1*(1-pi.tilde1) / (sum(c.table[1,])+2) +
  pi.tilde2*(1-pi.tilde2) / (sum(c.table[2,])+2)
agc1 <- pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.AC)
agc1
```

```
## [1] -0.02432022  0.01807142
```

Therefore 95% Agresti-Caffo confidence interval is -0.0243 $< \pi_1 - \pi_2 <$ 0.0181. Since this interval include zero, there is no sufficient evidence to indicate the difference in the proportions of male and female customers who got churned. Since both the confidence intervals at 90% confidence, failed to prove any difference in the proportion of male and female customers who got churned, we will proceed with the hypothesis test to determine whether to include this variable in the regression model. The hypothesis is given below:

$$H_o : \pi_{female} - \pi_{male} = 0$$
$$H_a : \pi_{female} - \pi_{male} \neq 0$$

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = c.table, conf.level = 0.90, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c.table
```

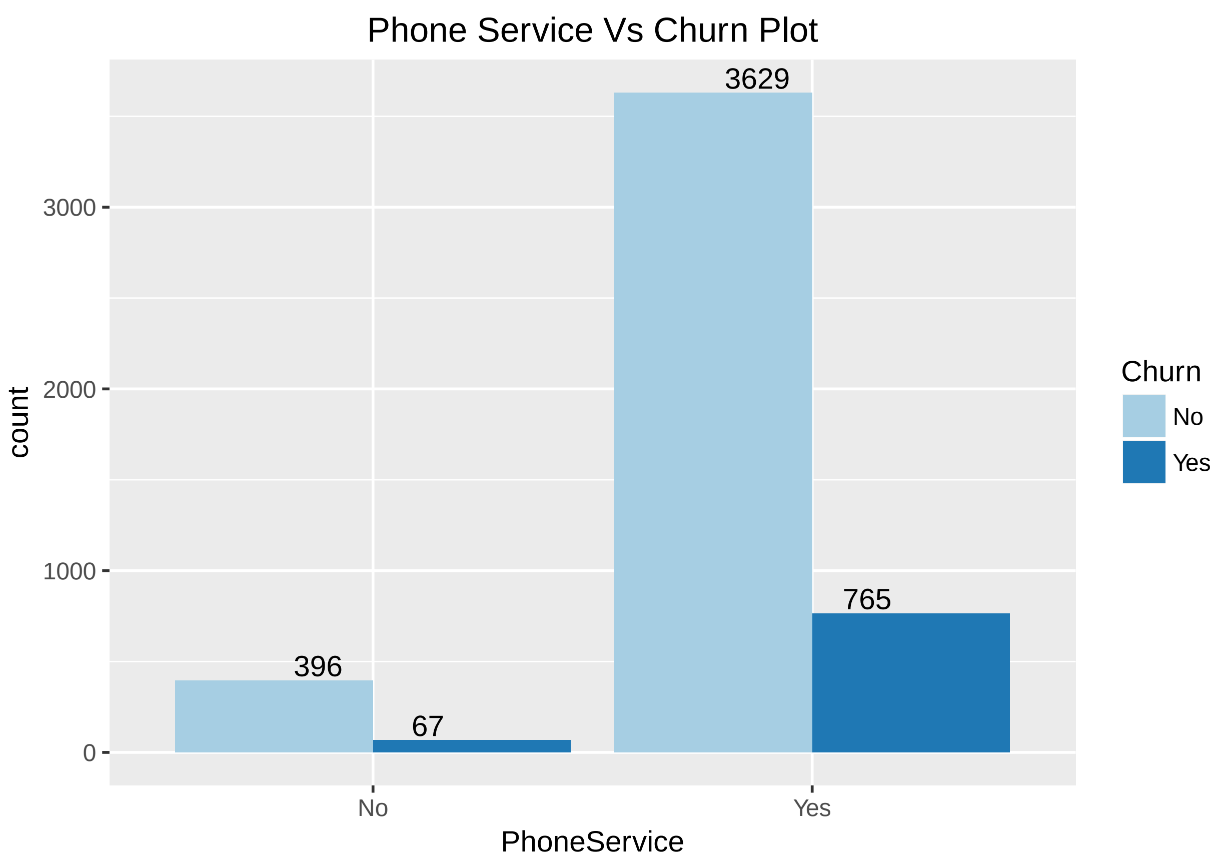
```
## X-squared = 0.083922, df = 1, p-value = 0.7721
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.02091671  0.01465185
## sample estimates:
##    prop 1    prop 2
## 0.1697171 0.1728496
```

Note that the the p-value = 0.7721 and $X^2 = Z_0^2 = 0.083922$. Root of $Z_0 = \sqrt{0.083922} = 0.289693$. Since $-1.645 < 0.289693 < 1.645$ we failed to reject H_0 when $\alpha = 0.1$. Also the wald confidence interval at 90% confidence level (-0.02091671 < 0 < 0.01465185) include zero. There is not sufficient evidence to conclude that the probability of female getting churned is different than probability of male getting churned. So we remove the variable from the analysis.

4.2.2 Phone Service

We have noticed from the Phone Service vs churn bar plot from the phase I that the ratio of churned and non-churned customers are similar for the customers using the phone service and those customers not using the phone service. The plot is given below.

```
ggplot(telcom_churn, aes(PhoneService,fill=Churn)) +
geom_bar(stat="count", position = "dodge") +
scale_fill_brewer(palette="Paired") +
ggtitle("Phone Service Vs Churn Plot") +
theme(plot.title = element_text(hjust = 0.5)) +
geom_text(aes(label=..count..),
stat="count",position=position_dodge(0.5),vjust=-0.2)
```



If there is not much difference in probability getting churned between the customers using the phone service and those customers not using the phone service, this variable can hardly provide any information about the churn tendency of the customers. A contigency table is constructed for the variables PhoneService and churn.

```r
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$PhoneService=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$PhoneService=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$PhoneService=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$PhoneService=="No")

ps.table<-array(data = c(YY,YN,NY,NN),
                dim = c(2,2),
                dimnames = list(PhoneService = c("Yes", "No"),
                Churn = c("Yes", "No")))
ps.table
```

```
##             Churn
## PhoneService Yes   No
##          Yes 765 3629
##          No   67  396
```

```r
ifelse(sum(ps.table)==length(telcom_churn$Churn),
       "Contingency table includes all the observations",
       "Contigency Table Made is not correct")
```

```
## [1] "Contingency table includes all the observations"
```

```r
# Find the estimated pi^j
pi.hat.table.ps<-ps.table/rowSums(ps.table)
pi.hat.table.ps
```

```
##             Churn
## PhoneService       Yes        No
##          Yes 0.1741010 0.8258990
##          No  0.1447084 0.8552916
```

17.41 % of customers with phone servide got churned and 14.47 % customers without phone service got churned. While 82.59 % phone service customers are still active, 85.53 % of the customers without phone service are active with Telco.

So now we have to check if there is a difference between the proportion of with and without phone service who got churned. For that purpose we will consider both Wald and Agresti-Caffo confidence intervals.

```r
# Confidence interval for difference of two probabilities
alpha<-0.05
pi.hat1.ps<-pi.hat.table.ps[1,1] # Proportion of customers using phone service who got churned
pi.hat2.ps<-pi.hat.table[2,1] # Proportion of customers not using phone service who got churned

# Wald CI
var.wald<-pi.hat1.ps*(1-pi.hat1.ps) / sum(ps.table[1,]) + pi.hat2.ps*(1-pi.hat2.ps) / sum(ps.table[2,])
wal1 <- pi.hat1.ps - pi.hat2.ps + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald)
wal1
```

```
## [1] -0.03496918  0.03747213
```

Therefore 95% Wald confidence interval is -0.035 $< \pi_1 - \pi_2 <$ 0.0375. Since this interval include zero, there is no sufficient evidence to indicate the difference in the proportions of churned customers using phone service and churned customers who are not using phone service.

```r
# Agresti-Caffo CI
pi.tilde1<-(ps.table[1,1]+1)/(sum(ps.table[1,])+2)
pi.tilde2<-(ps.table[2,1]+1)/(sum(ps.table[2,])+2)
var.AC<-pi.tilde1*(1-pi.tilde1) / (sum(ps.table[1,])+2) +
  pi.tilde2*(1-pi.tilde2) / (sum(ps.table[2,])+2)
agc1 <- pi.tilde1 - pi.tilde2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.AC)
agc1
```

```
## [1] -0.006004274  0.062029791
```

Therefore 95% Agresti-Caffo confidence interval is -0.006 $< \pi_1 - \pi_2 <$ 0.062. Since this interval include zero, there is no sufficient evidence to indicate the difference in the proportions of churned customers using phone service and churned customers who are not using phone service. Since both the confidence intervals failed to prove any difference in the proportion of both types of customers, we will proceed with the hypothesis test with 90 % confidence limit to determine whether to include this variable in the regression model. The hypothesis is given below:

$$H_o : \pi_{Phone} - \pi_{NoPhone} = 0$$
$$H_a : \pi_{Phone} - \pi_{NoPhone} \neq 0$$

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = ps.table, conf.level = 0.90, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  ps.table
## X-squared = 2.5492, df = 1, p-value = 0.1104
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  0.0009009569 0.0578842902
## sample estimates:
##    prop 1    prop 2
## 0.1741010 0.1447084
```

Note that the the p-value = 0.1104 and $X^2 = Z_0^2 = 2.5492$. Root of $Z_0 = \sqrt{2.5492} = 1.596621$. Since $-1.645 < 1.596621 < 1.645$ we failed to reject H_0 when $\alpha = 0.1$. Eventhough the wald confidence interval at 90% confidence level (0.0009009569 < w < 0.0578842902) does not include zero, its lower limit is almost equal to zero even at 90% confidence limit. There is not sufficient evidence to conclude that the probability of customers with phone service getting churned is different than probability of customers without phone service getting churned. So we remove the variable from the analysis.

4.3 Relative Risk

Since difference in probabilities measures a quantity whose meaning changes according to the sizes of $\pi_1 - \pi_2$, we would be interested in finding the relative risk of the variables by taking the ratio of two success probabilities.

```r
# Gender
cat("The sample relative risk is", round(pi.hat1/pi.hat2, 4), "\n \n")
```

```
## The sample relative risk is 0.9819
##
```

A customer getting churned is 0.9819 times as likely for females than for males.

```r
# Phone Service
cat("The sample relative risk is", round(pi.hat1.ps/pi.hat2.ps, 4), "\n \n")
```

```
## The sample relative risk is 1.0072
##
```

The probability of a customer getting churned is 1.0072 times as large for those customers using phone service than those who are not.

4.3.1 Confidence interval

The 95% confidence interval for the Relative risk for gender in customer churn in found as follows:

```r
# Gender
alpha<-0.05
n1<-sum(c.table[1,])
n2<-sum(c.table[2,])

# Wald confidence interval for RR of gender and churn
ci<-exp(log(pi.hat1/pi.hat2) + qnorm(p = c(alpha/2, 1-alpha/2)) *
          sqrt((1-pi.hat1)/(n1*pi.hat1) + (1-pi.hat2)/(n2*pi.hat2)))
round(ci, 4)  # relative risk for gender and churn
```

```
## [1] 0.8676 1.1112
```

```r
rev(round(1/ci, 4))  # inverted relative risk for gender and churn
```

```
## [1] 0.8999 1.1526
```

Since the relative risk of the gender and churn include 1, we confirm our analysis that gender is not suffieciently explaining the variability in churn tendency of a customer.

```r
# Phone Service
alpha<-0.05
n1<-sum(ps.table[1,])
n2<-sum(ps.table[2,])

# Wald confidence interval for RR of Phone service and churn
ci.ps<-exp(log(pi.hat1.ps/pi.hat2.ps) + qnorm(p = c(alpha/2, 1-alpha/2)) *
          sqrt((1-pi.hat1.ps)/(n1*pi.hat1.ps) + (1-pi.hat2.ps)/(n2*pi.hat2.ps)))
round(ci.ps, 4)  # relative risk for Phone service and churn
```

```
## [1] 0.8169 1.2419
```

```r
rev(round(1/ci.ps, 4))  # inverted relative risk for Phone service and churn
```

```
## [1] 0.8052 1.2241
```

Since the relative risk of the phone service and churn include 1, we confirm our analysis that phone service is not suffieciently explaining the variability in churn tendency of a customer.

4.4 Odds Ratio

By calculating the odds ratio which is the probability of success by probability of failure, we can estimate how large is the odds of a customer getting churned for different groups. If we take the case of Senior citizens, we can calculate how large is the odds of a senior citizen getting churned compared non-senior citizen. For that purpose we have to create a contingency table including variables 'SeniorCitizens' and 'churn'.

```r
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$SeniorCitizen=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$SeniorCitizen=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$SeniorCitizen=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$SeniorCitizen=="No")

sc.table<-array(data = c(YY,YN,NY,NN),
                dim = c(2,2),
                dimnames = list(SeniorCitizen = c("Yes", "No"),
                Churn = c("Yes", "No")))
sc.table
```

```
##              Churn
## SeniorCitizen Yes   No
##           Yes 259  563
##           No  573 3462
```

```r
ifelse(sum(sc.table)==length(telcom_churn$Churn),
       "Contingency table includes all the observations",
       "Contigency Table Made is not correct")
```

```
## [1] "Contingency table includes all the observations"
```

```r
# Find the estimated pi^j
pi.hat.table.sc<-sc.table/rowSums(sc.table)
pi.hat.table.sc
```

```
##              Churn
## SeniorCitizen       Yes        No
##           Yes 0.3150852 0.6849148
##           No  0.1420074 0.8579926
```

16.97 % of senior citizen got churned and 17.28 % non-senior citizen got churned. While 83.03 % senior citizens are still active customers, 82.72 of the non senior citizen are active with Telco.

```r
# Odds Ratio (OR)
alpha <- 0.05
OR.hat<-sc.table[1,1]*sc.table[2,2] / (sc.table[2,1]*sc.table[1,2])
round(OR.hat, 2)
```

```
## [1] 2.78
```

```r
round(1/OR.hat, 2) # Inverse of OR
```

```
## [1] 0.36
```

This result can be interpresed as the estimated odds of a customer getting churned are 2.78 times as large as for senior citizens than for non-senior citizens. It can also be said that the estimated odds of a customer getting churned are 0.36 times as large as in case of non-senior citizens than in case of senior citizens.

4.4.1 Confidence initerval

The confidence interval of the odds ratio is caculated as follows.

```r
var.log.or<-1/sc.table[1,1] + 1/sc.table[1,2] + 1/sc.table[2,1] + 1/sc.table[2,2]
OR.CI<-exp(log(OR.hat) + qnorm(p = c(alpha/2, 1-alpha/2)) *
             sqrt(var.log.or))
round(OR.CI, 2)
```

```
## [1] 2.34 3.30
```

```r
rev(round(1/OR.CI, 2))
```

```
## [1] 0.30 0.43
```

The 95% confidence interval for OR is 2.34 < OR < 3.3 If the interval is inverted, the 95% confidence interval for 1/OR is $0.3 < 1/OR < 0.43$.

4.5 Test of independence for multinomial variables

There are many multinomial variables in the dataset which needs to be analysed for their independence with each other. Most of these variables seems to have rebundant information as atleast one of their levels depends on the some other levels of a different variable. For that purpose we need to create a multinomial contigency table as shown below for the variables `PhoneService` and `MultipleLines`.

```r
# Multiple lines and phone service
levels(telcom_churn$MultipleLines)
```

```
## [1] "No"               "No phone service" "Yes"
```

```r
YY <- sum(telcom_churn$PhoneService=="Yes" & telcom_churn$MultipleLines=="Yes")
YN <- sum(telcom_churn$PhoneService=="Yes" & telcom_churn$MultipleLines=="No")
YPS <- sum(telcom_churn$PhoneService=="Yes" & telcom_churn$MultipleLines=="No phone service")
NY <- sum(telcom_churn$PhoneService=="No" & telcom_churn$MultipleLines=="Yes")
NN <- sum(telcom_churn$PhoneService=="No" & telcom_churn$MultipleLines=="No")
NPS <- sum(telcom_churn$PhoneService=="No" & telcom_churn$MultipleLines=="No phone service")

multi.table1<-array(data = c(YY, NY, YN, NN, YPS, NPS),
               dim = c(2,3),
               dimnames = list(PhoneService = c("Yes", "No"),
               MultipleLines = c("Yes", "No","No_P.S")))
multi.table1
```

```
##             MultipleLines
## PhoneService  Yes   No No_P.S
##          Yes 2472 1922      0
##          No     0    0    463
```

```r
ifelse(sum(multi.table1)==length(telcom_churn$MultipleLines),
       "Contingency table includes all the observations",
       "Contigency Table Made is not correct")
```

```
## [1] "Contingency table includes all the observations"
```

```r
chisq <- chisq.test(x = multi.table1, correct = FALSE)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  multi.table1
## X-squared = 4857, df = 2, p-value < 2.2e-16
```

Note the value $X^2 = 4857$ and p-value using $X^2 =$ is less than 2.2e-16. Because the p-value is extremely small, we reject the null hypothesis stating that values are independent.

So we see that the test of independence failed because of one category leaking information in to another. The information contained in the variable `PhoneService` and `MultipleLines` are rebundant. So we may have to remove the variable PhoneService.

Similarly we check the independence for other variables as well.

```
# Multiple lines and phone service
levels(telcom_churn$InternetService)
```

```
## [1] "DSL"         "Fiber optic" "No"
```

```
levels(telcom_churn$OnlineSecurity)
```

```
## [1] "No"                  "No internet service" "Yes"
```

```
DY <- sum(telcom_churn$InternetService=="DSL" & telcom_churn$OnlineSecurity=="Yes")
DN <- sum(telcom_churn$InternetService=="DSL" & telcom_churn$OnlineSecurity=="No")
DNS <- sum(telcom_churn$InternetService=="DSL" & telcom_churn$OnlineSecurity=="No internet service")
NY <- sum(telcom_churn$InternetService=="No" & telcom_churn$OnlineSecurity=="Yes")
NN <- sum(telcom_churn$InternetService=="No" & telcom_churn$OnlineSecurity=="No")
NNS <- sum(telcom_churn$InternetService=="No" & telcom_churn$OnlineSecurity=="No internet service")
FY <- sum(telcom_churn$InternetService=="Fiber optic" & telcom_churn$OnlineSecurity=="Yes")
FN <- sum(telcom_churn$InternetService=="Fiber optic" & telcom_churn$OnlineSecurity=="No")
FNS <- sum(telcom_churn$InternetService=="Fiber optic" & telcom_churn$OnlineSecurity=="No internet serv

multi.table2<-array(data = c(DY,NY,FY,DN,NN,FN,DNS,NNS,FNS),
               dim = c(3,3),
               dimnames = list(InternetService = c("DSL", "No", "Fiber optic" ),
               OnlineSecurity = c("Yes", "No","No_I.S")))
multi.table2
```

```
##                OnlineSecurity
## InternetService Yes   No No_I.S
##     DSL         995  676      0
##     No            0    0   1013
##     Fiber optic 753 1420      0
```

```
ifelse(sum(multi.table2)==length(telcom_churn$OnlineSecurity),
       "Contingency table includes all the observations",
       "Contigency Table Made is not correct")
```

```
## [1] "Contingency table includes all the observations"
```

```
chisq <- chisq.test(x = multi.table2, correct = FALSE)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  multi.table2
## X-squared = 5155.3, df = 4, p-value < 2.2e-16
```

Note the value $X^2 = 5155.2724125$ and p-value using $X^2 =$ is less than 2.2e-16. Because the p-value is extremely small, we reject the null hypothesis stating that values are independent.

So we see that the test of independence failed because of one category leaking information in to another. The information contained in the variable InternetService and OnlineSecurity are rebundant. People without internet service will not have any features associated with it. This includes features like OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies. So we may have to merge these categories in these variables.

```r
# Merging Levels
levels(telcom_churn$OnlineSecurity) <- list(Yes="Yes", No=c("No internet service", "No"))
levels(telcom_churn$OnlineBackup) <- list(Yes="Yes", No=c("No internet service", "No"))
levels(telcom_churn$DeviceProtection) <- list(Yes="Yes", No=c("No internet service", "No"))
levels(telcom_churn$TechSupport) <- list(Yes="Yes", No=c("No internet service", "No"))
levels(telcom_churn$StreamingTV) <- list(Yes="Yes", No=c("No internet service", "No"))
levels(telcom_churn$StreamingMovies) <- list(Yes="Yes", No=c("No internet service", "No"))
summary(telcom_churn)
```

```
##       customerID      gender     SeniorCitizen Partner    Dependents
##  0011-IGKFF:   1   Female:2404   No :4035      No :2018   No :3191
##  0013-SMEOE:   1   Male  :2453   Yes: 822      Yes:2839   Yes:1666
##  0014-BMAQU:   1
##  0016-QLJIS:   1
##  0017-DINOC:   1
##  0017-IUDMW:   1
##  (Other)   :4851
##  tenure   PhoneService          MultipleLines     InternetService
##  2:1024   No : 463     No              :1922   DSL        :1671
##  3: 832   Yes:4394     No phone service: 463   Fiber optic:2173
##  4: 762                Yes             :2472   No         :1013
##  5: 832
##  6:1407
##
##
##  OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV
##  Yes:1748       Yes:2088     Yes:2093         Yes:1761    Yes:2216
##  No :3109       No :2769     No :2764         No :3096    No :2641
##
##
##
##
##
##  StreamingMovies           Contract    PaperlessBilling
##  Yes:2237        Month-to-month:1881   No :1972
##  No :2620        One year      :1349   Yes:2885
##                  Two year      :1627
##
##
##
##
##                    PaymentMethod  MonthlyCharges    TotalCharges
##  Bank transfer (automatic):1315   Min.   : 18.25   Min.   : 218.6
##  Credit card (automatic)  :1297   1st Qu.: 43.95   1st Qu.:1312.2
##  Electronic check         :1387   Median : 75.40   Median :2607.6
##  Mailed check             : 858   Mean   : 68.66   Mean   :3181.9
##                                   3rd Qu.: 94.65   3rd Qu.:4863.9
##                                   Max.   :118.75   Max.   :8684.8
##
##  Churn
##  No :4025
##  Yes: 832
##
##
```

```
##
##
##
```

Now all those variables are merged to form binary variables. We have to again create the contingency table for these binary variables to test the hypothesis on difference in probabilities. By this way we can determine if these variable explains the variation in the churn behaviour.

```
# OnlineSecurity
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$OnlineSecurity=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$OnlineSecurity=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$OnlineSecurity=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$OnlineSecurity=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(OnlineSecurity = c("Yes", "No"),
               Churn = c("Yes", "No")))
X.table
```

```
##                Churn
## OnlineSecurity Yes   No
##            Yes 203 1545
##            No  629 2480
```

```
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 58.544, df = 1, p-value = 1.988e-14
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.10679825 -0.06556802
## sample estimates:
##    prop 1    prop 2
## 0.1161327 0.2023159
```

P-value is less than 0.05 and the 95% wald confidence interval doesnot include zero. This means that we failed to reject null hypothesis that there is no difference between the probabilities. So this variable is explaining some of the variability in the response variable.

```
# OnlineBackup
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$OnlineBackup=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$OnlineBackup=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$OnlineBackup=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$OnlineBackup=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(OnlineBackup = c("Yes", "No"),
               Churn = c("Yes", "No")))
X.table
```

```
##              Churn
```

```
## OnlineBackup Yes   No
##          Yes 360 1728
##          No  472 2297
```

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 0.032055, df = 1, p-value = 0.8579
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01946172  0.02337201
## sample estimates:
##    prop 1    prop 2
## 0.1724138 0.1704586
```

Note that the the p-value = 0.8579 and $X^2 = Z_0^2 = 0.032055$. We failed to reject H_0 when $\alpha = 0.05$. The wald confidence interval at 95% confidence level include zero. There is not sufficient evidence to conclude that the probability of probabilities of these variables are different. So we remove the variable from the analysis.

```r
# DeviceProtection
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$DeviceProtection=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$DeviceProtection=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$DeviceProtection=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$DeviceProtection=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(DeviceProtection = c("Yes", "No"),
               Churn = c("Yes", "No")))
X.table
```

```
##                 Churn
## DeviceProtection Yes   No
##              Yes 361 1732
##              No  471 2293
```

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 0.036108, df = 1, p-value = 0.8493
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01933673  0.02348570
## sample estimates:
##    prop 1    prop 2
## 0.1724797 0.1704052
```

Note that the the p-value = 0.8493 and $X^2 = Z_0^2 = 0.036108$. We failed to reject H_0 when $\alpha = 0.05$. The wald confidence interval at 95% confidence level include zero. There is not sufficient evidence to conclude that the probability of probabilities of these variables are different. So we remove the variable from the analysis.

```
# TechSupport
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$TechSupport=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$TechSupport=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$TechSupport=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$TechSupport=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(TechSupport = c("Yes", "No"),
               Churn = c("Yes", "No")))
X.table
```

```
##            Churn
## TechSupport Yes   No
##         Yes 203 1558
##         No  629 2467
```

```
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 61.083, df = 1, p-value = 5.473e-15
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.10846530 -0.06731463
## sample estimates:
##    prop 1    prop 2
## 0.1152754 0.2031654
```

P-value is less than 0.05 and the 95% wald confidence interval doesnot include zero. This means that we failed to reject null hypothesis that there is no difference between the probabilities. So this variable is explaining some of the variability in the response variable.

```
# StreamingTV
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$StreamingTV=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$StreamingTV=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$StreamingTV=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$StreamingTV=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(StreamingTV = c("Yes", "No"),
               Churn = c("Yes", "No")))
X.table
```

```
##            Churn
## StreamingTV Yes   No
##         Yes 501 1715
##         No  331 2310
```

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 86.163, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.07923977 0.12226366
## sample estimates:
##    prop 1    prop 2
## 0.2260830 0.1253313
```

P-value is less than 0.05 and the 95% wald confidence interval doesnot include zero. This means that we failed to reject null hypothesis that there is no difference between the probabilities. So this variable is explaining some of the variability in the response variable.

```r
# StreamingMovies
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$StreamingMovies=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$StreamingMovies=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$StreamingMovies=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$StreamingMovies=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(StreamingMovies = c("Yes", "No"),
               Churn = c("Yes", "No")))
X.table
```

```
##                Churn
## StreamingMovies Yes   No
##             Yes 508 1729
##             No  324 2296
```

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 90.929, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.08197103 0.12488043
## sample estimates:
##    prop 1    prop 2
## 0.2270899 0.1236641
```

P-value is less than 0.05 and the 95% wald confidence interval doesnot include zero. This means that we failed to reject null hypothesis that there is no difference between the probabilities. So this variable is explaining some of the variability in the response variable.

Now we analyse the variable Total charges.

```r
# Correlation between Monthly charges + Tenure and Total charges
cor(telcom_churn$TotalCharges,telcom_churn$MonthlyCharges)
```

```
## [1] 0.7602904
```

```r
smpl <- data.frame(
  Tenure=telcom_churn$tenure,
  MonthlyCharges=telcom_churn$MonthlyCharges,
  TotalCharges=telcom_churn$TotalCharges,
  CalulatedCharges=as.numeric(telcom_churn$tenure)*12*telcom_churn$MonthlyCharges)

head(smpl,5)
```

```
##   Tenure MonthlyCharges TotalCharges CalulatedCharges
## 1      3          56.95      1889.50           1366.8
## 2      4          42.30      1840.75           1522.8
## 3      2          89.10      1949.40           1069.2
## 4      3         104.80      3046.05           2515.2
## 5      6          56.15      3487.95           3369.0
```

```r
cor(smpl$TotalCharges,smpl$CalulatedCharges)
```

```
## [1] 0.9902616
```

The correlation between the calculated fields produced from the variables `MonthlyCharges` and `Tenure` is extremely correlated tot the variable `TotalCharges`. So it happens that the Total charges is a calculated measure of Tenure and Monthly charges. This leaks duplicate information in to the dataset. So we remove this variable from the analysis.

```r
# Partner and dependents
YY <- sum(telcom_churn$Partner=="Yes" & telcom_churn$Dependents=="Yes")
YN <- sum(telcom_churn$Partner=="Yes" & telcom_churn$Dependents=="No")
NY <- sum(telcom_churn$Partner=="No" & telcom_churn$Dependents=="Yes")
NN <- sum(telcom_churn$Partner=="No" & telcom_churn$Dependents=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(Dependents = c("Yes", "No"),
               Partner = c("Yes", "No")))
X.table
```

```
##           Partner
## Dependents  Yes   No
##        Yes 1449  217
##        No  1390 1801
```

```r
# C.I. and also hypothesis test for Ho: pi_1|1 - pi_1|2
prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 849.49, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
## 95 percent confidence interval:
##  0.4105430 0.4577525
## sample estimates:
##    prop 1    prop 2
## 0.8697479 0.4356001
```

P-value is less than 0.05 and the 95% wald confidence interval doesnot include zero. This means that we failed to reject null hypothesis that there is no difference between the probabilities. So this variable `Partner` is explaining some of the variability in the variable `Dependents`. This means that the variables may be giving rebundant information.

```
# Partner and dependents
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$Dependents=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$Dependents=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$Dependents=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$Dependents=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(Dependents = c("Yes", "No"),
               Churn = c("Yes", "No")))

prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 74.216, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.11850961 -0.07770776
## sample estimates:
##    prop 1    prop 2
## 0.1068427 0.2049514
```

```
# Partner and dependents
YY <- sum(telcom_churn$Churn=="Yes" & telcom_churn$Partner=="Yes")
YN <- sum(telcom_churn$Churn=="Yes" & telcom_churn$Partner=="No")
NY <- sum(telcom_churn$Churn=="No" & telcom_churn$Partner=="Yes")
NN <- sum(telcom_churn$Churn=="No" & telcom_churn$Partner=="No")

X.table<-array(data = c(YY,YN,NY,NN),
               dim = c(2,2),
               dimnames = list(Partner = c("Yes", "No"),
               Churn = c("Yes", "No")))

prop.test(x = X.table, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  X.table
## X-squared = 22.455, df = 1, p-value = 2.151e-06
```

```
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.07386417 -0.03010431
## sample estimates:
##    prop 1    prop 2
## 0.1497006 0.2016848
```

p-value of the variable `dependents` is much less than that of `partner`. Also wlad confidence interval is more close to zero for the variable `partner`. This may means that we are more confident is stating that the variable `dependents` explain some of the variability in churn than stating the same for `partner`.

Since there is only one numerical variable in the model, it need not to be standardised.

5 Logistic Regression

Sampling the data in to training and testing.

```r
smp_size <- floor(0.75 * nrow(telcom_churn))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(telcom_churn)), size = smp_size)

train <- telcom_churn[train_ind, ] # Training data set
test <- telcom_churn[-train_ind, ] # Testing data set
```

5.1 Model Building

The model is buid using the Genaralised Linear Model with family as binomial and link as logit. All of those variables we have finalised for the model after initial analysis are included in to the model.

```r
mod.fit1<-glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines + InternetService +
              family = binomial(link = logit), data = train)

summary(mod.fit1) # Summary of the model
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges, family = binomial(link = logit), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6817  -0.5751  -0.2838  -0.1313   3.1983
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           0.24806    1.17131   0.212 0.832276
## SeniorCitizenYes                      0.22770    0.12036   1.892 0.058509
## DependentsYes                        -0.14336    0.12320  -1.164 0.244555
## tenure3                              -0.47951    0.14473  -3.313 0.000923
```

```
## tenure4                              -0.54174    0.16258  -3.332 0.000862
## tenure5                              -0.70933    0.16949  -4.185 2.85e-05
## tenure6                              -1.01678    0.20526  -4.954 7.29e-07
## MultipleLinesNo phone service        -0.33870    0.37396  -0.906 0.365084
## MultipleLinesYes                      0.46128    0.13987   3.298 0.000974
## InternetServiceFiber optic            1.33510    0.39323   3.395 0.000686
## InternetServiceNo                    -1.74606    0.51849  -3.368 0.000758
## OnlineSecurityNo                      0.11787    0.13886   0.849 0.396001
## TechSupportNo                         0.13944    0.14264   0.978 0.328303
## StreamingTVNo                        -0.50173    0.19473  -2.577 0.009980
## StreamingMoviesNo                    -0.50376    0.19135  -2.633 0.008472
## ContractOne year                     -0.64535    0.13937  -4.630 3.65e-06
## ContractTwo year                     -1.52977    0.22878  -6.687 2.28e-11
## PaperlessBillingYes                   0.30179    0.12308   2.452 0.014207
## PaymentMethodCredit card (automatic)  0.05357    0.15770   0.340 0.734106
## PaymentMethodElectronic check         0.55250    0.13482   4.098 4.17e-05
## PaymentMethodMailed check            -0.10601    0.20519  -0.517 0.605407
## MonthlyCharges                       -0.02429    0.01452  -1.673 0.094341
##
## (Intercept)
## SeniorCitizenYes                     .
## DependentsYes
## tenure3                              ***
## tenure4                              ***
## tenure5                              ***
## tenure6                              ***
## MultipleLinesNo phone service
## MultipleLinesYes                     ***
## InternetServiceFiber optic           ***
## InternetServiceNo                    ***
## OnlineSecurityNo
## TechSupportNo
## StreamingTVNo                        **
## StreamingMoviesNo                    **
## ContractOne year                     ***
## ContractTwo year                     ***
## PaperlessBillingYes                  *
## PaymentMethodCredit card (automatic)
## PaymentMethodElectronic check        ***
## PaymentMethodMailed check
## MonthlyCharges                       .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3294.7  on 3641  degrees of freedom
## Residual deviance: 2483.9  on 3620  degrees of freedom
## AIC: 2527.9
##
## Number of Fisher Scoring iterations: 6
```

There are many coefficients for the model which are not significant. The function stepAIC from the mass package is used for variable selection. It is an iterative process in which variables are added and removed, in order to get a subset of variables that gives the best performing model.

```
mod.fit2<- stepAIC(mod.fit1, direction="both")

summary(mod.fit2)
```

It is seen that many insignificant variables have been removed from the model. After various trial and error methods, the following model is considered as the final model for the evaluation. Trying to remove the variable `MonthlyCharges` also increased the significance of many other variables. But when this is used as an interaction term in the model, model is performing much better. Interaction is found between variables `MultipleLines` and `MonthlyCharges`. It makes sense as more the number of lines a customer is having more will be his monthly charges. The model is defined below.

```
mod.fit3<-glm(formula = Churn ~ SeniorCitizen + tenure + InternetService +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
     MultipleLines:MonthlyCharges,
              family = binomial(link = logit), data = train)

summary(mod.fit3)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + tenure + InternetService +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     MultipleLines:MonthlyCharges, family = binomial(link = logit),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6315  -0.5973  -0.2968  -0.1271   3.1594
##
## Coefficients:
##                                               Estimate Std. Error z value
## (Intercept)                                    1.70571    0.75933   2.246
## SeniorCitizenYes                               0.30133    0.11635   2.590
## tenure3                                       -0.47374    0.14278  -3.318
## tenure4                                       -0.52337    0.16016  -3.268
## tenure5                                       -0.69614    0.16701  -4.168
## tenure6                                       -1.04334    0.20242  -5.154
## InternetServiceFiber optic                     1.77249    0.27421   6.464
## InternetServiceNo                             -2.16528    0.44480  -4.868
## StreamingTVNo                                 -0.66791    0.16021  -4.169
## StreamingMoviesNo                             -0.68064    0.15877  -4.287
## ContractOne year                              -0.74873    0.13716  -5.459
## ContractTwo year                              -1.69697    0.22481  -7.548
## PaperlessBillingYes                            0.34939    0.12174   2.870
## MultipleLinesNo:MonthlyCharges                -0.03964    0.01016  -3.901
## MultipleLinesNo phone service:MonthlyCharges -0.05373    0.01583  -3.394
## MultipleLinesYes:MonthlyCharges               -0.03355    0.00958  -3.502
##                                               Pr(>|z|)
## (Intercept)                                   0.024683 *
## SeniorCitizenYes                              0.009602 **
## tenure3                                       0.000907 ***
## tenure4                                       0.001084 **
## tenure5                                       3.07e-05 ***
```

```
## tenure6                                        2.54e-07 ***
## InternetServiceFiber optic                     1.02e-10 ***
## InternetServiceNo                              1.13e-06 ***
## StreamingTVNo                                  3.06e-05 ***
## StreamingMoviesNo                              1.81e-05 ***
## ContractOne year                               4.79e-08 ***
## ContractTwo year                               4.41e-14 ***
## PaperlessBillingYes                            0.004105 **
## MultipleLinesNo:MonthlyCharges                 9.58e-05 ***
## MultipleLinesNo phone service:MonthlyCharges 0.000690 ***
## MultipleLinesYes:MonthlyCharges                0.000462 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3294.7  on 3641  degrees of freedom
## Residual deviance: 2516.0  on 3626  degrees of freedom
## AIC: 2548
##
## Number of Fisher Scoring iterations: 6
```

Now all the coefficients are significant at 95% confidence level. Other than the coefficient `SeniorCitizenYes` all other coefficients are significant atleast 99% confidence level. AIC (Akaike's Information Criteria) is high with AIC:2548, which is not that great sign.

5.2 Hypothesis Tests

```r
round(summary(mod.fit3)$coefficients, 4)    # Wald tests
```

```
##                                              Estimate Std. Error z value
## (Intercept)                                    1.7057     0.7593  2.2463
## SeniorCitizenYes                               0.3013     0.1163  2.5899
## tenure3                                       -0.4737     0.1428 -3.3180
## tenure4                                       -0.5234     0.1602 -3.2677
## tenure5                                       -0.6961     0.1670 -4.1683
## tenure6                                       -1.0433     0.2024 -5.1545
## InternetServiceFiber optic                     1.7725     0.2742  6.4640
## InternetServiceNo                             -2.1653     0.4448 -4.8680
## StreamingTVNo                                 -0.6679     0.1602 -4.1690
## StreamingMoviesNo                             -0.6806     0.1588 -4.2870
## ContractOne year                              -0.7487     0.1372 -5.4590
## ContractTwo year                              -1.6970     0.2248 -7.5483
## PaperlessBillingYes                            0.3494     0.1217  2.8700
## MultipleLinesNo:MonthlyCharges                -0.0396     0.0102 -3.9009
## MultipleLinesNo phone service:MonthlyCharges  -0.0537     0.0158 -3.3935
## MultipleLinesYes:MonthlyCharges               -0.0335     0.0096 -3.5018
##                                              Pr(>|z|)
## (Intercept)                                    0.0247
## SeniorCitizenYes                               0.0096
## tenure3                                        0.0009
## tenure4                                        0.0011
## tenure5                                        0.0000
```

```
## tenure6                                        0.0000
## InternetServiceFiber optic                     0.0000
## InternetServiceNo                              0.0000
## StreamingTVNo                                  0.0000
## StreamingMoviesNo                              0.0000
## ContractOne year                               0.0000
## ContractTwo year                               0.0000
## PaperlessBillingYes                            0.0041
## MultipleLinesNo:MonthlyCharges                 0.0001
## MultipleLinesNo phone service:MonthlyCharges   0.0007
## MultipleLinesYes:MonthlyCharges                0.0005
```

```r
anova(mod.fit3, test = "Chisq")  # Sequential testing of variables
```

```
## 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                                         3641     3294.7
## SeniorCitizen                1    94.63      3640     3200.1 < 2.2e-16
## tenure                       4   190.70      3636     3009.4 < 2.2e-16
## InternetService              2   341.61      3634     2667.8 < 2.2e-16
## StreamingTV                  1     7.76      3633     2660.0  0.005332
## StreamingMovies              1     3.08      3632     2656.9  0.079066
## Contract                     2   109.20      3630     2547.7 < 2.2e-16
## PaperlessBilling             1     9.32      3629     2538.4  0.002271
## MultipleLines:MonthlyCharges  3    22.39      3626     2516.0 5.416e-05
##
## NULL
## SeniorCitizen                ***
## tenure                       ***
## InternetService              ***
## StreamingTV                  **
## StreamingMovies              .
## Contract                     ***
## PaperlessBilling             **
## MultipleLines:MonthlyCharges ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod.fit3, mod.fit1, test = "Chisq") #comparing two models
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ SeniorCitizen + tenure + InternetService + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + MultipleLines:MonthlyCharges
## Model 2: Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
```

```
##     MonthlyCharges
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3626     2516.0
## 2      3620     2483.9  6   32.135 1.538e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis of Deviance Table shows the amount by which the model `mod.fit3` deviates from the model `mod.fit1`. The deviance is given as 21.342. Model `mod.fit3` deviates from the observed data by 2494.7 and model `mod.fit1` deviates from the observed data by 2516.0.

6 Model evaluation

6.1 Goodness of fit (GOF)

Residual measures are obtained for the model to check how well a model fits on the individual observations.

```
# Goodness-of-Fit Tests
rdev <- mod.fit3$deviance  # deviance
rdev
```

```
## [1] 2516.044
```

```
dfr <- mod.fit3$df.residual # degree of freedom
dfr
```

```
## [1] 3626
```

```
ddf <- rdev/dfr # for a reasonable model this should not be far from 1
ddf
```

```
## [1] 0.6938898
```

```
thresh2 <- 1 + 2*sqrt(2/dfr)
thresh3 <- 1 + 3*sqrt(2/dfr)
c(thresh2, thresh3)
```

```
## [1] 1.046971 1.070457
```

The deviance for the model `mod.fit3` is given as 2516.0444663 and degree of freedom is given as 3626. The rario of deviance and the degree of freedom is used to measure the goodness of the fit for the model, which is given as 0.6938898. To check if this is too far from 1 for the model created, we check this using the threshold values, 1.0469711, 1.0704567. Since $0.6938898<1.0469711$ and also $0.6938898<1.0704567$ the model is a good fit.

7 Prediction

The model was then tested on the testing data and the predicted results were compared with the observed data.

```
linear.pred<-predict(object = mod.fit3, newdata = test, type = "link")
pred <- data.frame(Real=test$Churn, predicted=ifelse(linear.pred<0.5, "No", "Yes"))
confusionMatrix(pred$Real, pred$predicted)
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction  No Yes
##        No  992   2
##        Yes 213   8
##
##                Accuracy : 0.823
##                  95% CI : (0.8004, 0.8441)
##     No Information Rate : 0.9918
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0544
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8232
##             Specificity : 0.8000
##          Pos Pred Value : 0.9980
##          Neg Pred Value : 0.0362
##              Prevalence : 0.9918
##          Detection Rate : 0.8165
##    Detection Prevalence : 0.8181
##       Balanced Accuracy : 0.8116
##
##        'Positive' Class : No
##
```

The model was able to explain the 82.3% variability in the response variable. The 95% confidence interval of the model is given as 0.8004 and 0.8441. The positive class was taken as customers who have not churned as those are the customers we are interested in.

$$\begin{aligned} Sensitivity &: 0.8232 \\ Specificity &: 0.8000 \\ PositivePredValue &: 0.9980 \\ NegativePredValue &: 0.0362 \\ Prevalence &: 0.9918 \\ DetectionRate &: 0.8165 \\ BalancedAccuracy &: 0.8116 \end{aligned}$$

8 Results and Discussion

It was found that there were many variable with inter dependencies. Some variables were rebundant and were leaking duplicate information in to the model. Such variables were identified and removed. Variables like InternetService and PhoneService were already included in many other variables. Also it was found that some binary variables like PhoneService and Gender were not able to explain the variance in the response variable. Such variables were identified and removed. The variable SeniorCitizen was having the maximum effect on the response variable. It was able explain much of the variance in the response variable. Contigency tables were constructed for analysis of independence. Some multinomial variable were found to be independ of the response variable. The model was build using the most important identified variables. After many trial and error iterations, we came up with the final variable that also included an interaction term between multiple lines and monthly charges. It makes sense as more the number of lines a customer is having more will be his monthly charges. The model was evaluated using the goodness of the fit test. The deviance of the model was found to be 2516.0444663 and the ratio between the deviation and the degree of freedom was close to one, which proved that the model is well fitting the data. The prediction results showed that the model performance was satisfactory. 82.3 % of the variablity in the response variable was explained by the model. The sensitivity of the model was also very high reaching 82.32%.

9 Conclusion

The data had several issues including interdependency and duplicate information leaking in to other variables. The other issues identified included lack of independece, rebundant information and high correlation between some calculated fields. The model was build using the variables that were identified to be relevant for explaining the variance in the response variable. Various statistical techniques including cinfidence intervals, hypothesis analysis, tests for independence, Odds ratio were used at multi times in the analysis to get the best model. The model was satisfactory in predicting the churned and non churned customer in test dataset.