Business Data Mining

IDS 572, Spring 2020

Team Members

Anuj Chanchlani (674153696)

Pratik Talreja (657488876)

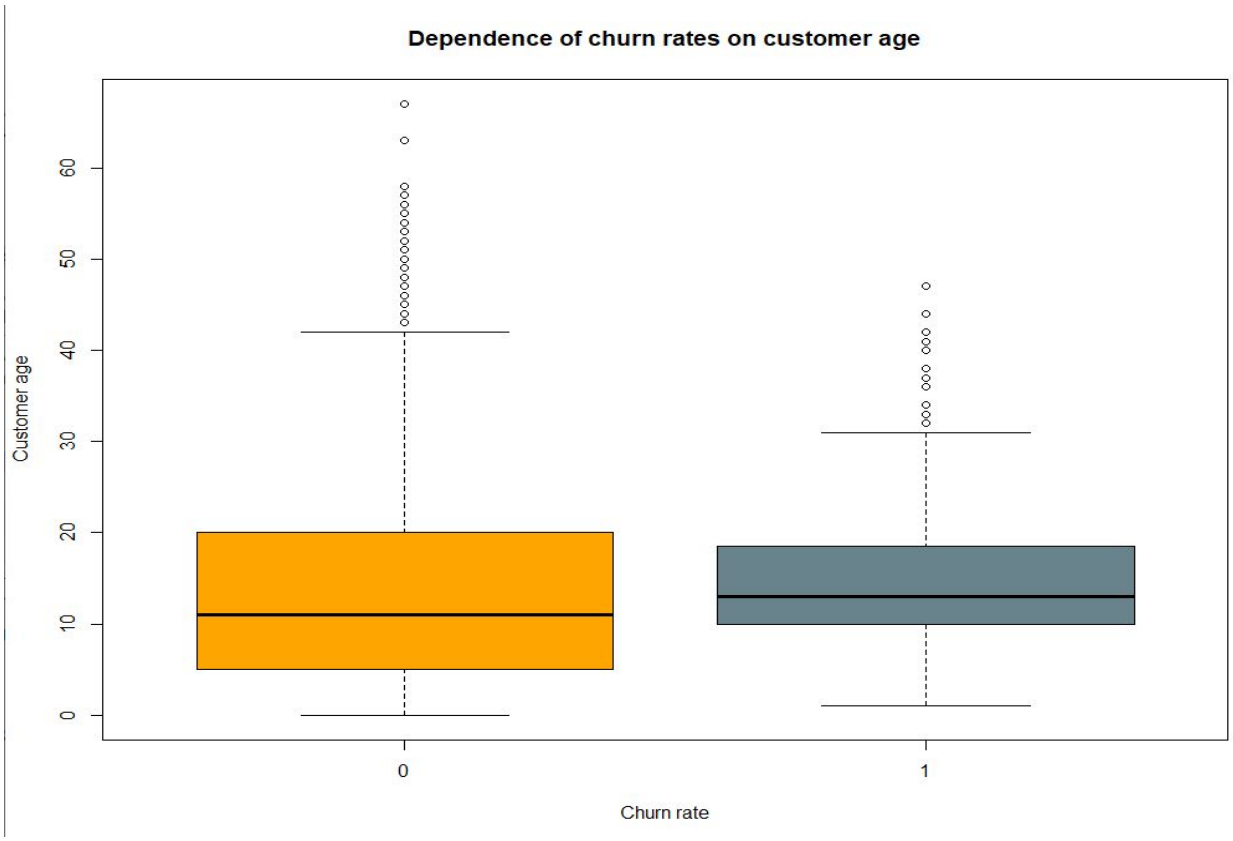
Rashi Desai (663553314)

```
# Importing required packages and libraries
install.packages("readxl")
library(readxl)
#Loading the dataset
data = read_excel("dataset_4.xlsx")
#Churn rate originally being a numeric field has to be converted to factor using #as.factor
data$`Churn (1 = Yes, 0 = No)`= as.factor(data$`Churn (1 = Yes, 0 = No)`)
```

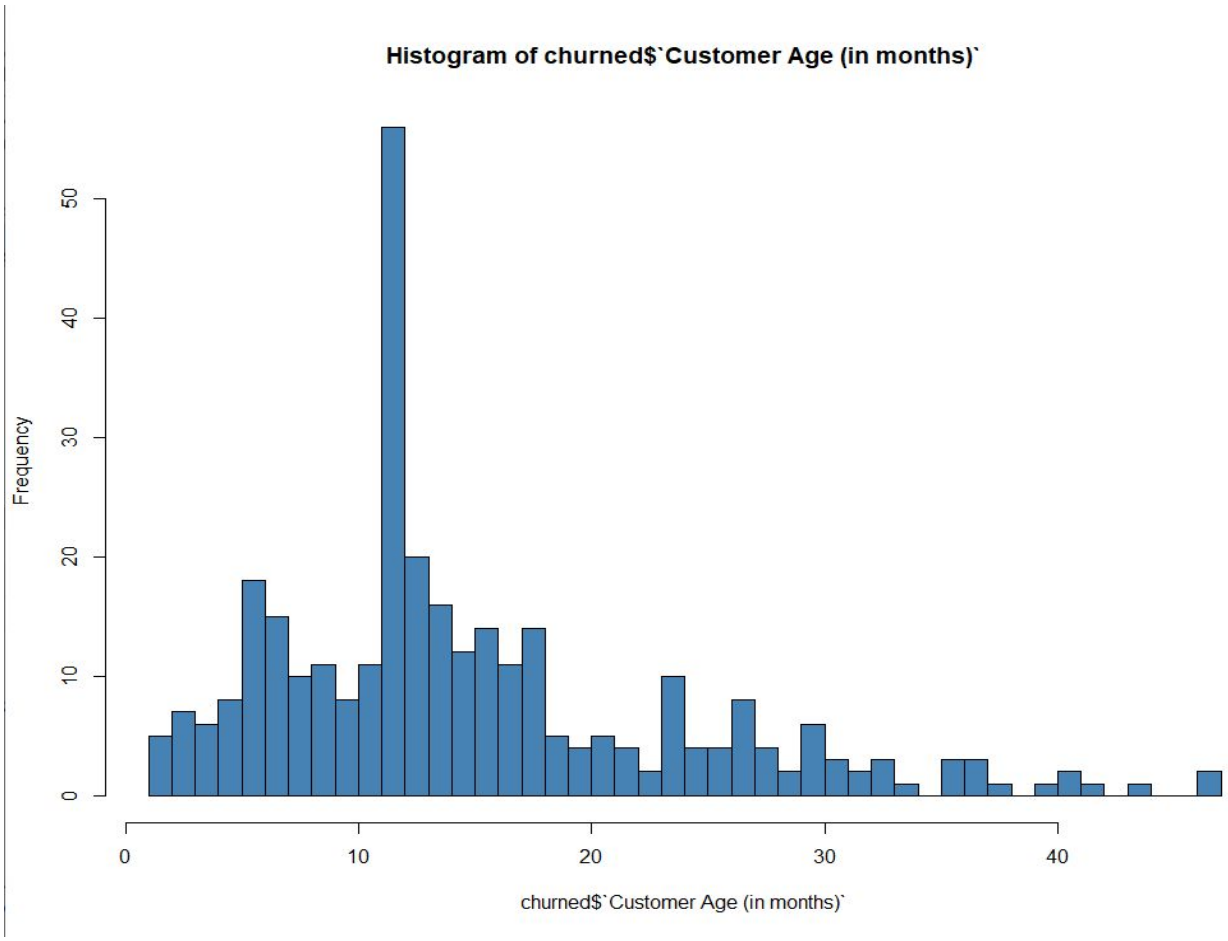
#Q1

#To determine if there exists a relation between two variables, we can plot the variables against each other.
Customer Age - *numeric*
Churn rate - *factor*

```
boxplot(data$`Customer Age (in months)` ~ data$`Churn (1 = Yes, 0 = No)`, data=data, main="Dependence of churn rates on customer age", xlab="Churn rate", ylab="Customer age",col=c("orange", "lightblue4"))
```



# now extracting only the churned customers and plotting it against their respective #ages.
churned <- data[data$`Churn (1 = Yes, 0 = No)` == 1,]
hist(churned$`Customer Age (in months)`, col = ("steelblue"),probability = FALSE,breaks = 50)



- From the box plot, we observe that the median age (in months) for customers who left was 13 months and the customers who never left was 11 months.
- By observing the histogram, we see that between intervals 6-14 months, the frequency increases and reaches a peak when the customer age is 12 months. This means that in the case of churned customers, the large number of them chose to unsubscribe after utilizing the services for 12 months.
- After the period of service utilization goes beyond 14 months, we see a decrease in the number of churned customers strengthening Walls belief ”customers utilizing QWE’s services for more than 14 months knows how to use them and therefore are less likely to leave”
- From our analysis, we can conclude that there is less churning for young and old aged customers. But those between 6-14 months are the riskiest group.

- From the above visualizations, we can conclude that there exists a dependence of churn rates on customer age.

 # Q2

To best predict the probability of customer churning for records 672, 354 and 5203, we first extract those records from the dataset.

```
record_672 <- data[data$ID==672,]
record_354 <- data[data$ID==354,]
record_5203 <- data[data$ID==5203,]

data <- data[-c(672, 354, 5203),]

# split the dataset into training and testing data sets.

set.seed(256)
index <- sample(2, nrow(data), replace = T ,  prob = c(0.8,0.2))
train <- data[index == 1,]
test <- data[index == 2,]

model_1 <- glm(`Churn (1 = Yes, 0 = No)` ~ . , data = train, family = "binomial")
summary(model_1)
```

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -9.899e-01  2.301e-01  -4.301 1.70e-05 ***
ID                         -3.796e-04  4.459e-05  -8.511  < 2e-16 ***
`Customer Age (in months)` -2.801e-02  7.461e-03  -3.755 0.000174 ***
`CHI Score Month 0`        -6.269e-03  1.384e-03  -4.529 5.93e-06 ***
`CHI Score 0-1`            -1.088e-02  2.813e-03  -3.868 0.000110 ***
`Support Cases Month 0`    -1.190e-01  1.132e-01  -1.051 0.293062
`Support Cases 0-1`         1.403e-01  9.835e-02   1.426 0.153784
`SP Month 0`                1.613e-03  1.122e-01   0.014 0.988529
`SP 0-1`                   -9.477e-03  8.743e-02  -0.108 0.913676
`Logins 0-1`                1.651e-03  2.111e-03   0.782 0.434256
`Blog Articles 0-1`        -1.007e-03  2.398e-02  -0.042 0.966512
`Views 0-1`                -2.851e-05  6.538e-05  -0.436 0.662825
`Days Since Last Login 0-1` 1.483e-02  4.454e-03   3.330 0.000869 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.9  on 5047  degrees of freedom
Residual deviance: 1882.6  on 5035  degrees of freedom
AIC: 1908.6

Number of Fisher Scoring iterations: 7
```

#From the above screenshot, we can that variables Customer age (in months), Chi #Score Month 0, Chi Score 0-1 and Days Since Last Login 0-1 are highly significant.
removing ID and other non-significant variables.

# Running the regression model now on test data
ans <- predict(model_1, newdata = test)

model_2 <- glm(`Churn (1 = Yes, 0 = No)`~ `Customer Age (in months)`+`CHI Score Month 0`+`CHI Score 0-1`+`Days Since Last Login 0-1`,data = train, family = "binomial")
summary(model_2)

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.795613   0.119265 -23.440  < 2e-16 ***
`Customer Age (in months)`  0.008190   0.005929   1.381 0.167224
`CHI Score Month 0`      -0.004342   0.001185  -3.663 0.000249 ***
`CHI Score 0-1`          -0.012365   0.002521  -4.906 9.31e-07 ***
`Days Since Last Login 0-1`  0.021505   0.005038   4.268 1.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.9  on 5047  degrees of freedom
Residual deviance: 1959.0  on 5043  degrees of freedom
AIC: 1969

Number of Fisher Scoring iterations: 6
```

From the screenshot, we can see that the variable Customer Age suddenly becomes non-significant.

But from Q1, we know that there lies some dependence between Customer Age and Churn.

Now let's categorize the Customer Age variable into three categories '0-5','6-14'
 # and 14-.
dummy <- data
dummy$cat[data$`Customer Age (in months)`>=0 & data$`Customer Age (in months)`<6] <- '0-6'
dummy$cat[data$`Customer Age (in months)`>=6 & data$`Customer Age (in months)`<=14] <- '6-14'
dummy$cat[data$`Customer Age (in months)`>14] <- '14-'

```
# Converting it into factor.
dummy$cat <-as.factor(dummy$cat)

# Removing column ID and Customer Age in months
dummy <- dummy[,-1] # Execute this particular line of code twice.

set.seed(256)
index <- sample(2, nrow(dummy), replace = T ,  prob = c(0.8,0.2))
train_1 <- dummy[index == 1,]
test_1 <- dummy[index == 2,]

model_3 <- glm(`Churn (1 = Yes, 0 = No)`~ `CHI Score Month 0`+`CHI Score 0-1`+`Days Since Last Login 0-1`+ cat ,data = train_1, family = "binomial")
summary(model_3)
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.846303   0.221219 -17.387  < 2e-16 ***
`CHI Score Month 0`      -0.009171   0.001281  -7.159 8.13e-13 ***
`CHI Score 0-1`          -0.006753   0.002588  -2.609  0.00907 **
`Days Since Last Login 0-1`  0.010204   0.003963   2.574  0.01004 *
cat14-                    1.731594   0.263138   6.581 4.69e-11 ***
cat6-14                   2.053715   0.253123   8.114 4.92e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.9  on 5047  degrees of freedom
Residual deviance: 1874.7  on 5042  degrees of freedom
AIC: 1886.7
```

```
# As we moved from model1 to model2 the AIC increased and also Customer Age became insignificant.
# But by categorizing the Customer Age variable we were able to achieve significance for the Customer Age and also AIC improved.

# We will predict the probabilities of the test set and customers using 354, 672 and # 5203 customer this model
# We see that it returns positive values and negative values for the instances.
# These are nothing but predicted log of the odds.
ans <- predict(model_3, newdata = test_1)

# to get probabilities write (type = response)
ans <- predict(model_3, newdata = test_1, type = "response")
```

```
# to predict the classes based on probabilities
class <- ifelse(ans>=0.5,"YES","NO")
answer <- data.frame(class)

#------------
```

Now, let's check probabilities of customer **672**

```
test1 <- record_672
test1$cat <- '14-'  # Assigning the category based on the age
test1$cat <- as.factor(test1$cat)
test1 <- test1[,-1] # executing it twice
predict(model_3,newdata = test1 , type = "response")
```

probability that the customer 672 will leave is 0.0305. It is low.

By looking at the dataset we come to know that between December 2011 and February 2012, the customer will **never leave**.

Now, let's check probabilities of customer **354**

```
test2 <- record_354
test2$cat <-'6-14'  # Assigning the category based on the age
test2$cat <- as.factor(test2$cat)
test2 <- test2[,-1] # executing it twice
predict(model_3,newdata = test1 , type = "response")
```

probability that the customer 354 will leave is 0.0530. It is low.

By looking at the dataset we come to know that between December 2011 and February 2012, the customer **never left.**

Now, let's check probabilities of customer **5203**

```
test3 <- record_5203
test3$cat <-'0-6'  # Assigning the category based on the age
test3$cat <- as.factor(test3$cat)
test3 <- test3[,-1] # executing it twice
predict(model_3,newdata = test3 , type = "response")
```

probability that the customer 5203 will leave is 0.0127. It is low.

By looking at the dataset we come to know that between December 2011 and February 2012, the customer **never left.**

#Q3

```
# Creating the data frame of predicted probabilities
```

```
probabilities <- data.frame(ans)

# arranging the records in the decreasing order
probabilities<- data.frame(probabilities[with(probabilities, order(-ans)), ])

# extracting the top 100 records.
top_100 <- probabilities[1:100,]

#Now for determining top drivers for each of the 100 customers, we execute the
#following lines of code

caret::varImp(model_3)
```

	Overall
`CHI Score Month 0`	7.158975
`CHI Score 0-1`	2.609454
`Days Since Last Login 0-1`	2.574479
cat14-	6.580552
cat6-14	8.113523

```
summary(model_3)
```

Based on the snapshot and above line, we conclude that **CHI Score Month 0, CHI Score 0-1 and Customer Age (in months)** are the top drivers for the customers in the QWE churn data.