~~~
---
title: "Customer Churn Analysis"
output: html_notebook
---

```{r}
# Load the desired packages
library(tidyverse)
library(caTools)
library(rms)
```

```{r}
# Load the data
data <- read_csv('Customer-Churn.csv')
glimpse(data)
```

```{r}
# Look for missing data
lapply(data, function(x) unique(is.na(x)))
```

```{r}
# Substitute the mean for the missing values
data$TotalCharges <- ifelse(is.na(data$TotalCharges),
                            mean(data$TotalCharges, na.rm = TRUE),
                            data$TotalCharges)

# Check the outcome
unique(is.na(data$TotalCharges))
summary(data)
glimpse(data)
```

```{r}
# Prepare for Logistic Regression
# Check the levels in the discrete data.
lapply(data, function(x) unique(x))
```

```{r}
# The levels are inconsistent across the variables.  Rework the
variables so that the levels across the discrete data are consistent.
data$gender <- factor(data$gender,
                      levels = c('Female', 'Male'),
                      labels = c(1, 2))

data$SeniorCitizen <- factor(data$SeniorCitizen,
                      levels = c(0, 1),
~~~

```
                      labels = c(1, 2))

data$Partner <- factor(data$Partner,
                       levels = c('No', 'Yes'),
                       labels = c(1, 2))

data$Dependents <- factor(data$Dependents,
                          levels = c('No', 'Yes'),
                          labels = c(1, 2))

data$PhoneService <- factor(data$PhoneService,
                            levels = c('No', 'Yes'),
                            labels = c(1, 2))

data$MultipleLines <- factor(data$MultipleLines,
                             levels = c('No', 'Yes', 'No phone
service'),
                             labels = c(1, 2, 3))

data$InternetService <- factor(data$InternetService,
                               levels = c('No', 'DSL', 'Fiber optic'),
                               labels = c(1, 2, 3))

data$OnlineSecurity <- factor(data$OnlineSecurity,
                              levels = c('No', 'Yes', 'No internet
service'),
                              labels = c(1, 2, 3))

data$OnlineBackup <- factor(data$OnlineBackup,
                            levels = c('No', 'Yes', 'No internet
service'),
                            labels = c(1, 2, 3))

data$DeviceProtection <- factor(data$DeviceProtection,
                            levels = c('No', 'Yes', 'No internet
service'),
                            labels = c(1, 2, 3))

data$TechSupport <- factor(data$TechSupport,
                            levels = c('No', 'Yes', 'No internet
service'),
                            labels = c(1, 2, 3))

data$StreamingTV <- factor(data$StreamingTV,
                            levels = c('No', 'Yes', 'No internet
service'),
                            labels = c(1, 2, 3))

data$StreamingMovies <- factor(data$StreamingMovies,
                            levels = c('No', 'Yes', 'No internet
service'),
```

~~~~
                            labels = c(1, 2, 3))

data$Contract <- factor(data$Contract,
                        levels = c('Month-to-month', 'One year', 'Two
year'),
                        labels = c(1, 2, 3))

data$PaperlessBilling <- factor(data$PaperlessBilling,
                                levels = c('No', 'Yes'),
                                labels = c(1, 2))

data$PaymentMethod <- factor(data$PaymentMethod,
                             levels = c('Credit card (automatic)', 'Bank
transfer (automatic)',
                                        'Electronic check', 'Mailed
check'),
                             labels = c(1, 2, 3, 4))

data$Churn <- as.integer(ifelse(data$Churn == 'Yes', 1, 0))
```

```{r}
# Visual analysis of attributes that may be affecting customer churn
# Dependents
data %>%
  group_by(Dependents) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = Dependents, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(name = 'Dependents',
                   breaks = c('1', '2'),
                   labels = c('No', 'Yes'))
```

```{r}
# Total Charges
data %>%
  mutate(customer_spend_level = case_when(TotalCharges < 1000 ~ 1,
                                    TotalCharges >= 1000 & TotalCharges
< 3000 ~ 2,
                                    TotalCharges >= 3000 & TotalCharges
< 4000 ~ 3,
                                    TotalCharges >= 4000 & TotalCharges
< 5000 ~ 4,
                                    TotalCharges >= 5000 & TotalCharges
< 6000 ~ 5,
                                    TotalCharges >= 6000 ~ 6)) %>%
  group_by(customer_spend_level) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = customer_spend_level, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
~~~~

~~~
  xlab('Customer Charges (spend level)')
```

```{r}
# Monthly Charges
data %>%
  mutate(monthly_spend_level = case_when(MonthlyCharges < 20 ~ 1,
                                    MonthlyCharges >= 20 &
MonthlyCharges < 50 ~ 2,
                                    MonthlyCharges >= 50 &
MonthlyCharges < 100 ~ 3,
                                    MonthlyCharges >= 100 ~ 4)) %>%
  group_by(monthly_spend_level) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = monthly_spend_level, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  xlab('Monthly Charges (spend level)')
```

```{r}
# Tenure
data %>%
  mutate(customer_loyalty_level = case_when(tenure < 12 ~ 1,
                                            tenure >= 12 & tenure < 24 ~
2,
                                            tenure >= 24 & tenure < 36 ~
3,
                                            tenure >= 36 & tenure < 48 ~
4,
                                            tenure >= 48 & tenure < 60 ~
5,
                                            tenure >= 60 ~ 6)) %>%
  group_by(customer_loyalty_level) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = customer_loyalty_level, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  xlab('Customer Tenure (loyalty level)')
```

```{r}
# Paperless Billing
data %>%
  group_by(PaperlessBilling) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = PaperlessBilling, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks = c('1', '2'),
        labels = c('No', 'Yes'))
```

```{r}
~~~

~~~
# Payment Method
data %>%
  group_by(PaymentMethod) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = PaymentMethod, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks = c('1', '2', '3', 4),
        labels = c('Credit Card', 'Bank Transfer', 'E-Check', 'M-
Check'))
```

```{r}
# Contract Type
data %>%
  group_by(Contract) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = Contract, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  xlab('Contract Type') +
  scale_x_discrete(breaks = c('1', '2', '3'),
        labels = c('Month to Month', 'One Year', 'Two Year'))
```

```{r}
# Internet Service Type
data %>%
  group_by(InternetService) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = InternetService, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  xlab('Internet Service Type') +
  scale_x_discrete(breaks = c('1', '2', '3'),
        labels = c('No', 'DSL', 'Fiber Optic'))
```

```{r}
# Senior Citizen
data %>%
  group_by(SeniorCitizen) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = SeniorCitizen, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks=c('1', '2'),
        labels=c('No', 'Yes'))
```

```{r}
# Multiple Lines
data %>%
  group_by(MultipleLines) %>%
  summarise(churn = sum(Churn)) %>%
~~~

~~~
  ggplot(aes(x = MultipleLines, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks=c('1', '2', '3'),
        labels=c('No', 'Yes', 'No Phone Service'))
```

```{r}
# Streaming TV
data %>%
  group_by(StreamingTV) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = StreamingTV, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks=c('1', '2', '3'),
        labels=c('No', 'Yes', 'No Internet Service'))
```

```{r}
# Streaming Movies
data %>%
  group_by(StreamingMovies) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = StreamingMovies, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks=c('1', '2', '3'),
        labels=c('No', 'Yes', 'No Internet Service'))
```

```{r}
# Tech Support
data %>%
  group_by(TechSupport) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = TechSupport, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks=c('1', '2', '3'),
        labels=c('No', 'Yes', 'No Internet Service'))
```

```{r}
# Online Security
data %>%
  group_by(OnlineSecurity) %>%
  summarise(churn = sum(Churn)) %>%
  ggplot(aes(x = OnlineSecurity, y = churn)) +
  geom_bar(stat = 'identity', aes(fill = churn), color = 'black') +
  scale_x_discrete(breaks=c('1', '2', '3'),
        labels=c('No', 'Yes', 'No Internet Service'))
```

```{r}
~~~

~~~
# Base code to obtain the base persona churn totals and ratio
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Analysis secondary attributes with base persona
# Payment Method
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 &
PaymentMethod == 3) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Dependents
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 & Dependents
== 1) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Partner
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 & Partner ==
1) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Monthly Charges
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 &
MonthlyCharges >= 50 &
         MonthlyCharges < 100) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Senior Citizen
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 &
SeniorCitizen == 1) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
~~~

~~~
/ n())
```

```{r}
# Tech Support
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 &
TechSupport == 1) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Online Security
data %>%
  filter(tenure < 12 & InternetService == 3 & Contract == 1 &
OnlineSecurity == 1) %>%
  summarise(total_customers = n(), churn = sum(Churn), pct_churn = churn
/ n())
```

```{r}
# Scale Monthly and Total Charges
data[, 19:20] <- scale(data[, 19:20])
```

```{r}
# Create 'test' and 'train' data sets
set.seed(1000)
split <- sample.split(data$Churn, SplitRatio = 0.75)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)
```

```{r}
# Create the Logistic Regression Model. P-Value threshold of 0.05 was
used as variable inclusion criteria.
classifier <- glm(Churn ~ SeniorCitizen +
                          Dependents +
                          tenure +
                          MultipleLines +
                          InternetService +
                          OnlineSecurity +
                          TechSupport +
                          StreamingTV +
                          StreamingMovies +
                          Contract +
                          PaperlessBilling +
                          PaymentMethod +
                          MonthlyCharges +
                          TotalCharges,
~~~

~~~
                          family = 'binomial', data = train)

summary(classifier)
```

```{r}
# Check to see if any of our variable are being artificially inflated
vif(classifier)
```

```{r}
# Use the model to predict Churn
prob_pred <- predict(classifier, type = 'response', newdata = test)
y_pred <- ifelse(prob_pred >= 0.5, 1, 0)

# Create Confusion Matrix to show the model's performance
cm <- table(test$Churn, y_pred)
cm
```
~~~