Capstone Project

Foundations of Data Science Satya Dalai

January 2016

Index

The Problem

Customer Churn is a measure of the number of customers who leave or stop a service within a defined period, generally switching to a competitor within the same industry. Besides helping understand market fit and customer metrics, it is important because the costs of customer acquisition outweigh customer retention.

My Capstone will focus on a Customer Churn dataset in the telecom industry (https://github.com/ericchiang/churn/blob/master/data/churn.csv)

Churn Prediction

Telecom companies maintain large amounts of information on their customers used extensively to study ways to predict and reduce churn. Studying past behavior can identify leading factors of churn and predict customers who are most likely to churn. Offering them incentivized offers can help mitigate customer churn.

The dataset available to me has various details on State Location, Call Durations, Calling Plans, Customer Service Calls and Churn Status. I have used relevant details in my analysis from the information given.; and Classification to predict whether a customer will churn or not based on available metrics.

The Dataset

The Churn dataset has 3333 observations with 21 attributes listed below.

1	State
2	Account Length
3	Area Code
4	Phone Number
5	International Plan
6	Voice Mail Plan
7	Voice Mail Message
8	Day Minutes

9	Day Calls
10	Day Charge
11	Evening Minutes
12	Evening Calls
13	Evening Charge
14	Night Minutes
15	Night Calls
16	Night Charge

17	International Minutes
18	International Calls
19	International Charges

20	Customer Service Calls
21	Churn

Table 1. Churn Dataset Attributes

I followed the steps below to conduct my analysis:

1. Cleaning the Data
2. Exploratory Data Analysis
3. Logistic Regression Modeling
4. Decision Trees
5. Summary.

Analysis

#Reading the Churn dataset from CSV File.

churn<-read.csv("churn.csv")

#Exploring the nature of dataset with structure/names and summary functions.
#Names of all attributes
names(churn)

```
 [1] "State"          "Account.Length" "Area.Code"      "Phone"          "Int.l.Plan"     "VMail.Plan"
 [7] "VMail.Message"  "Day.Mins"       "Day.Calls"      "Day.Charge"     "Eve.Mins"       "Eve.Calls"
[13] "Eve.Charge"     "Night.Mins"     "Night.Calls"    "Night.Charge"   "Intl.Mins"      "Intl.Calls"
[19] "Intl.Charge"    "CustServ.Calls" "Churn."
```

#The names appear to have periods within them that I will clean.

#Structure of Dataset
str(churn)

```
'data.frame':   3333 obs. of  21 variables:
 $ State         : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 25 19 50 ...
 $ Account.Length: int  128 107 137 84 75 118 121 147 117 141 ...
 $ Area.Code     : int  415 415 415 408 415 510 510 415 408 415 ...
 $ Phone         : Factor w/ 3333 levels "327-1058","327-1319",..: 1927 1576 1118 1708 111 2254 1048 81 292
118 ...
 $ Int.l.Plan    : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...
 $ VMail.Plan    : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...
 $ VMail.Message : int  25 26 0 0 0 0 24 0 0 37 ...
 $ Day.Mins      : num  265 162 243 299 167 ...
 $ Day.Calls     : int  110 123 114 71 113 98 88 79 97 84 ...
 $ Day.Charge    : num  45.1 27.5 41.4 50.9 28.3 ...
 $ Eve.Mins      : num  197.4 195.5 121.2 61.9 148.3 ...
 $ Eve.Calls     : int  99 103 110 88 122 101 108 94 80 111 ...
 $ Eve.Charge    : num  16.78 16.62 10.3 5.26 12.61 ...
 $ Night.Mins    : num  245 254 163 197 187 ...
 $ Night.Calls   : int  91 103 104 89 121 118 118 96 90 97 ...
 $ Night.Charge  : num  11.01 11.45 7.32 8.86 8.41 ...
 $ Intl.Mins     : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
 $ Intl.Calls    : int  3 3 5 7 3 6 7 6 4 5 ...
 $ Intl.Charge   : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
 $ CustServ.Calls: int  1 1 0 2 3 0 3 0 1 0 ...
 $ Churn.        : Factor w/ 2 levels "False.","True.": 1 1 1 1 1 1 1 1 1 1 ...
```

#Summary of Dataset

```
summary(churn)
```

```
    State         Account.Length    Area.Code          Phone      Int.l.Plan VMail.Plan VMail.Message
WV     : 106   Min.   :  1.0   Min.   :408.0   327-1058:   1   no :3010   no :2411   Min.   : 0.000
MN     :  84   1st Qu.: 74.0   1st Qu.:408.0   327-1319:   1   yes: 323   yes: 922   1st Qu.: 0.000
NY     :  83   Median :101.0   Median :415.0   327-3053:   1                         Median : 0.000
AL     :  80   Mean   :101.1   Mean   :437.2   327-3587:   1                         Mean   : 8.099
OH     :  78   3rd Qu.:127.0   3rd Qu.:510.0   327-3850:   1                         3rd Qu.:20.000
OR     :  78   Max.   :243.0   Max.   :510.0   327-3954:   1                         Max.   :51.000
(Other):2824                                   (Other) :3327
   Day.Mins       Day.Calls       Day.Charge       Eve.Mins       Eve.Calls       Eve.Charge
Min.   :  0.0   Min.   :  0.0   Min.   : 0.00   Min.   :  0.0   Min.   :  0.0   Min.   : 0.00
1st Qu.:143.7   1st Qu.: 87.0   1st Qu.:24.43   1st Qu.:166.6   1st Qu.: 87.0   1st Qu.:14.16
Median :179.4   Median :101.0   Median :30.50   Median :201.4   Median :100.0   Median :17.12
Mean   :179.8   Mean   :100.4   Mean   :30.56   Mean   :201.0   Mean   :100.1   Mean   :17.08
3rd Qu.:216.4   3rd Qu.:114.0   3rd Qu.:36.79   3rd Qu.:235.3   3rd Qu.:114.0   3rd Qu.:20.00
Max.   :350.8   Max.   :165.0   Max.   :59.64   Max.   :363.7   Max.   :170.0   Max.   :30.91

  Night.Mins     Night.Calls     Night.Charge      Intl.Mins       Intl.Calls      Intl.Charge
Min.   : 23.2   Min.   : 33.0   Min.   : 1.040   Min.   : 0.00   Min.   : 0.000   Min.   :0.000
1st Qu.:167.0   1st Qu.: 87.0   1st Qu.: 7.520   1st Qu.: 8.50   1st Qu.: 3.000   1st Qu.:2.300
Median :201.2   Median :100.0   Median : 9.050   Median :10.30   Median : 4.000   Median :2.780
Mean   :200.9   Mean   :100.1   Mean   : 9.039   Mean   :10.24   Mean   : 4.479   Mean   :2.765
3rd Qu.:235.3   3rd Qu.:113.0   3rd Qu.:10.590   3rd Qu.:12.10   3rd Qu.: 6.000   3rd Qu.:3.270
Max.   :395.0   Max.   :175.0   Max.   :17.770   Max.   :20.00   Max.   :20.000   Max.   :5.400

CustServ.Calls     Churn.
Min.   :0.000   False.:2850
1st Qu.:1.000   True. : 483
Median :1.000
Mean   :1.563
3rd Qu.:2.000
Max.   :9.000
```

#The summary function gives an overview of the dataset. I can see here that Intnl Plan, VMail Plan and Churn are category variables.

Cleaning Data

#Cleaning the periods in the dataset names

```
names(churn)<-gsub("\\.","",names(churn))
```

#Renaming category variables to binary digits to facilitate classification analysis.

```
churn$VMailPlan<-ifelse(churn$VMailPlan=="no",0,1)
churn$IntlPlan<-ifelse(churn$IntlPlan=="no",0,1)
churn$Churn<-as.integer(churn$Churn)
```

```
churn$Churn[churn$Churn=="1"]<-0
churn$Churn[churn$Churn=="2"]<-1

#Changing binary classes to factor variables
churn$Churn<-as.factor(churn$Churn)
churn$IntlPlan<-as.factor(churn$IntlPlan)
churn$VMailPlan<-as.factor(churn$VMailPlan)

#The data set has certain metrics by day, night and evening, and I want to see if the sums will influence churn. I summed up total minutes, calls and charges and also bucketed account lengths with an interval of 50.

churn$totalcalls<-churn$DayCalls+churn$EveCalls+churn$NightCalls
churn$totalmins<-churn$DayMins+churn$EveMins+churn$NightMins
churn$totalcharge<-churn$DayCharge
+churn$EveCharge
+churn$NightCharge
+churn$IntlCharge
churn$AccountBucket<-cut(churn$AccountLength,c(0,50,100,150,200,250))
```

Exploratory Data Analysis

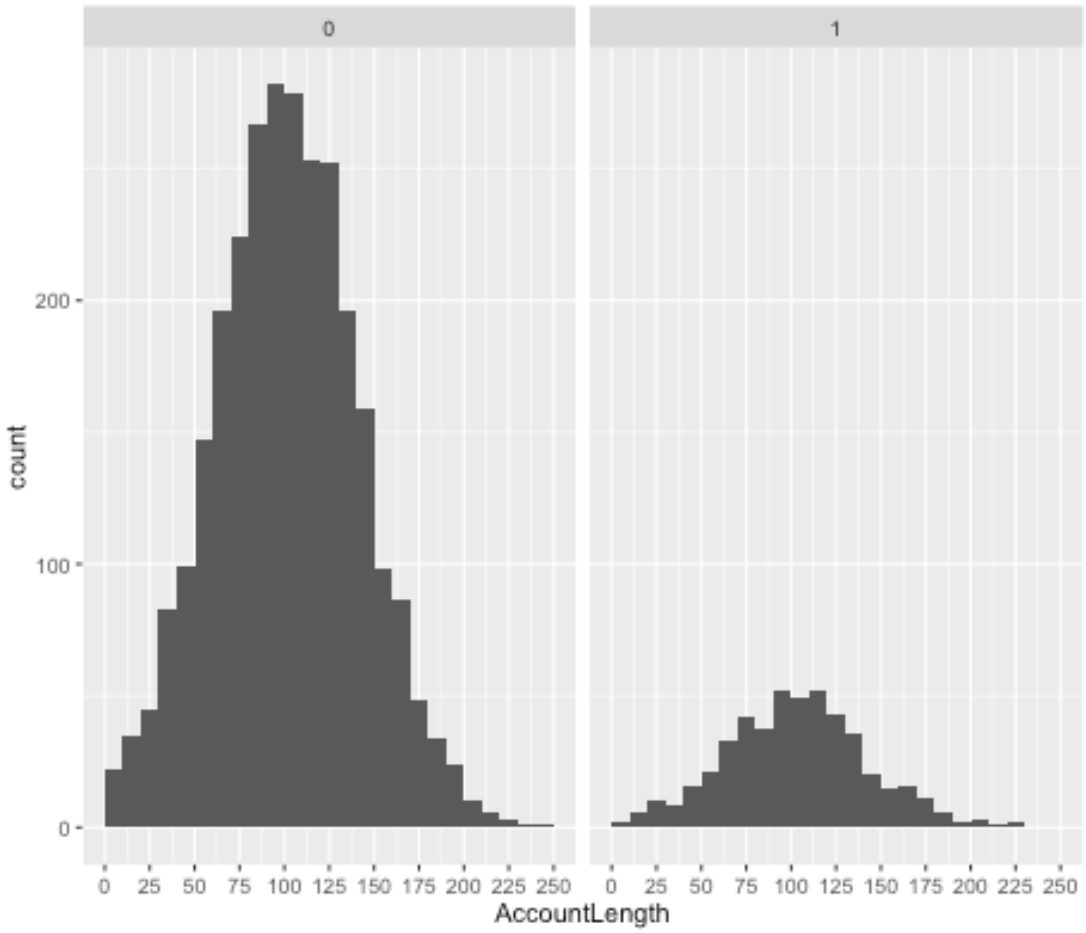
```
#Conducting EDA now to explore relationships between variables. Loading libraries dplyr and ggplot2
library(dplyr)
library(ggplot2)

#Looking at Churn Status here, we can see ~14.5% of accounts are churners.
table(churn$Churn)

False.  True.
 2850    483
```

#Looking at Churn by Account Length to ascertain whether tenure length with the service may have an effect on Churn.

```
qplot(x=AccountLength,data=subset(churn,!is.na(Churn)),binwidth=10)
        +scale_x_continuous(lim=c(0,250), breaks =seq(0,250,25))
        +facet_wrap(~Churn)
```



Churn by Account Length

#The majority of Accounts look to be in the same bracket of Account Length.
#Looking at the stats below, we can confirm this.

```
by(churn$AccountLength,churn$Churn, summary)
```

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    73.0   100.0   100.8   127.0   243.0
------------------------------------------------
churn$Churn: 1
```

#Looking at Churn by Voice Mail Messages.

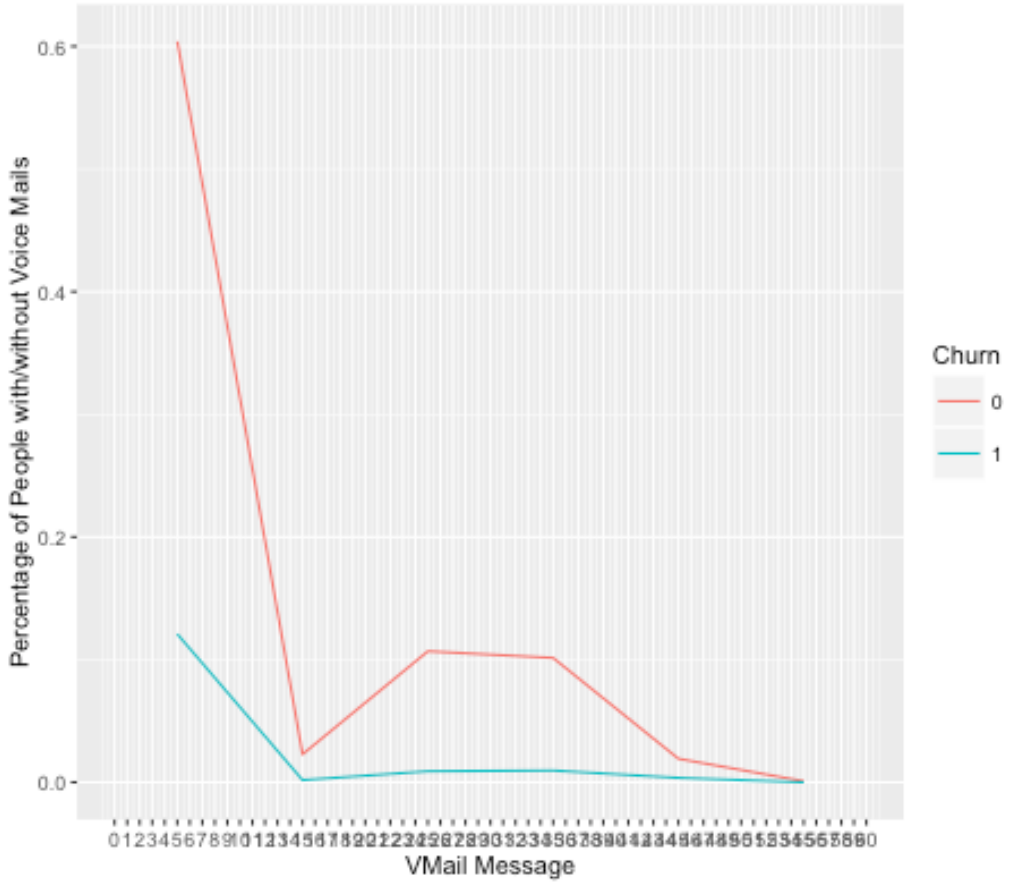
by(churn$VMailMessage,churn$Churn, summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   8.605  22.000  51.000
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   5.116   0.000  48.000
```

#50% of both churners and non churners have 0 voice mail messages. The frequency plot below shows the spread of both, which doesn't appear to have a pattern.

```
ggplot(aes(x = VMailMessage, y = ..count../sum(..count..)), data = subset(churn, !is.na(Churn)))
    +geom_freqpoly(aes(color = Churn), binwidth=10)
    +scale_x_continuous(limits = c(0, 60), breaks = seq(0, 60, 1))
    +xlab('VMail Message')
    +ylab('Percentage of People with/without Voice Mails')
```

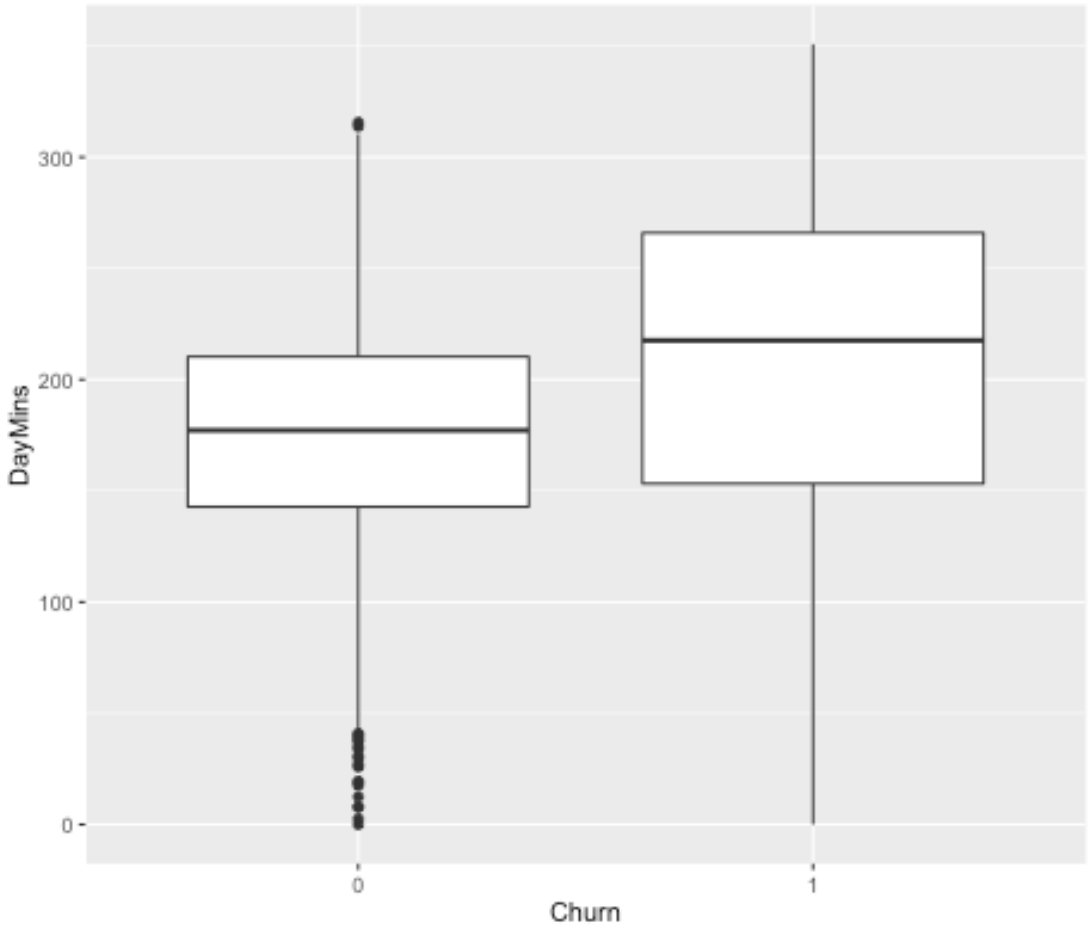
Percentage of Churners and Non Churners by Voice Mail Messages



#Churn by Day Minutes

#The boxplots below represent Churn by Day Minutes. The majority of churners seem to be heavier users than non churners. The Churn plot is bottom skewed and has a wider range while day minute usage is less variable in the non churners group.

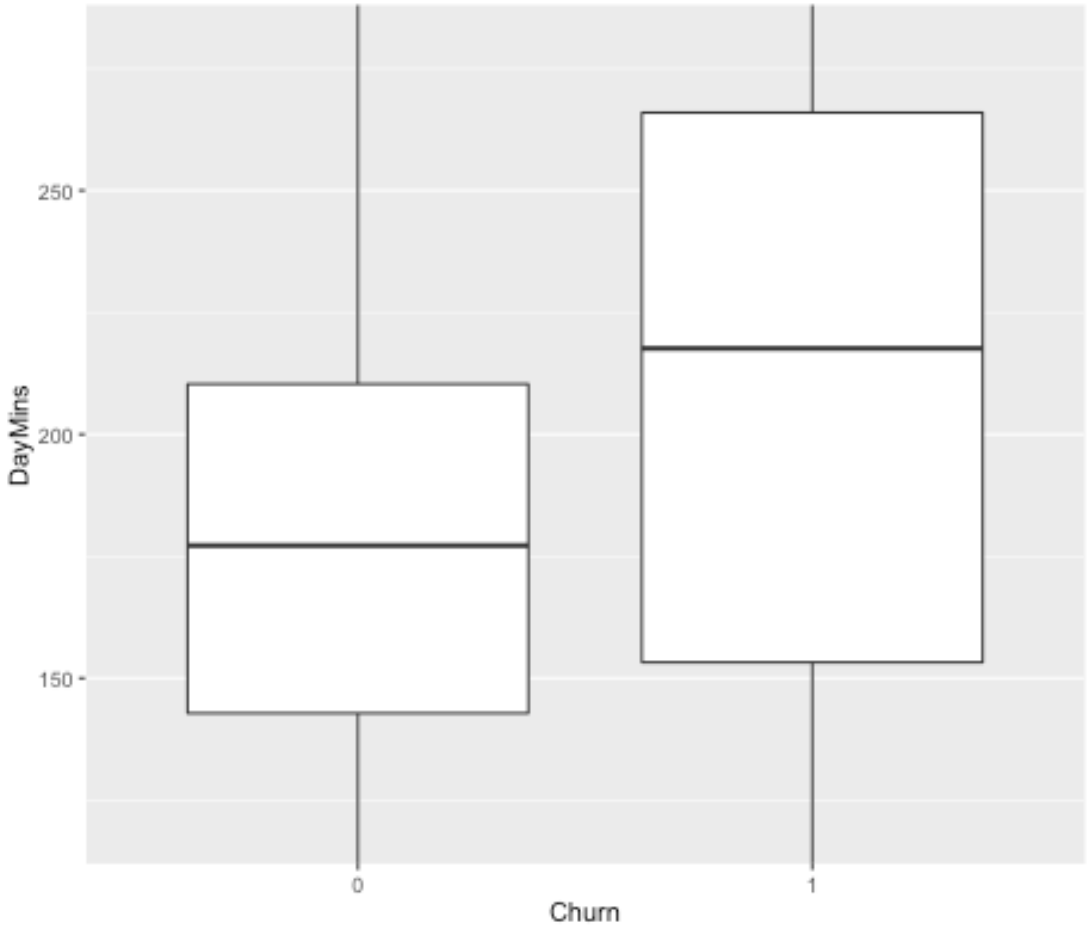
```
ggplot(aes(Churn,DayMins),data=subset(churn,!is.na(Churn)))+geom_boxplot()
```



Churn by Day Minutes

#Lets limit the Y axis and take a closer look

```
ggplot(aes(Churn,DayMins),data=subset(churn,!is.na(Churn)))+geom_boxplot()
+coord_cartesian(ylim=c(120,280))
```



Limited Y Axis - Churn by Day Minutes

#75% Non Churners are almost equal to the bottom half of Churners as we can see above. The 3rd quartile for non churners is ~210, close to the median ~218 for Churners.

by(churn$DayMins,churn$Churn, summary)

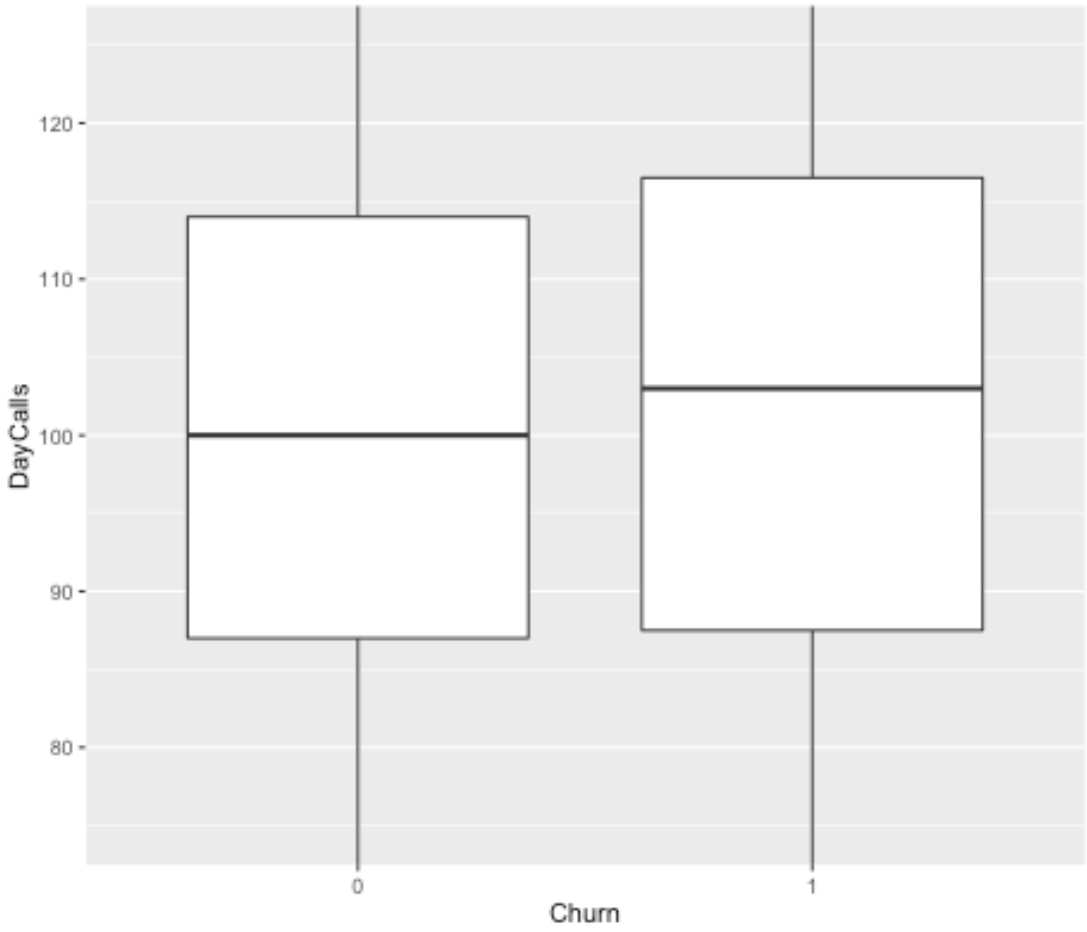
```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   142.8   177.2   175.2   210.3   315.6
------------------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   153.2   217.6   206.9   266.0   350.8
```

The higher day minutes attributed to Churners could possibly mean that frequent users of the telecom service encountered more issues increasing their churn rate.

#Churn by Day Calls

#The boxplots below represent Churn by Day Calls.

```
ggplot(aes(Churn,DayCalls),data=subset(churn,!is.na(Churn)))
        +geom_boxplot()
        +coord_cartesian(ylim=c(75,125))
```



Churn by Day Calls

#The boxplot doesn't appear to show a distinctive difference. Lets take a closer look.

```
by(churn$DayCalls,churn$Churn, summary)
```

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    87.0   100.0   100.3   114.0   163.0
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    87.5   103.0   101.3   116.5   165.0
```

As the plots showed, there doesn't appear to be a marked difference between churners and non churners. However, we know that Churners use more minutes than non churners, and so although they're making more calls, their calls last longer.

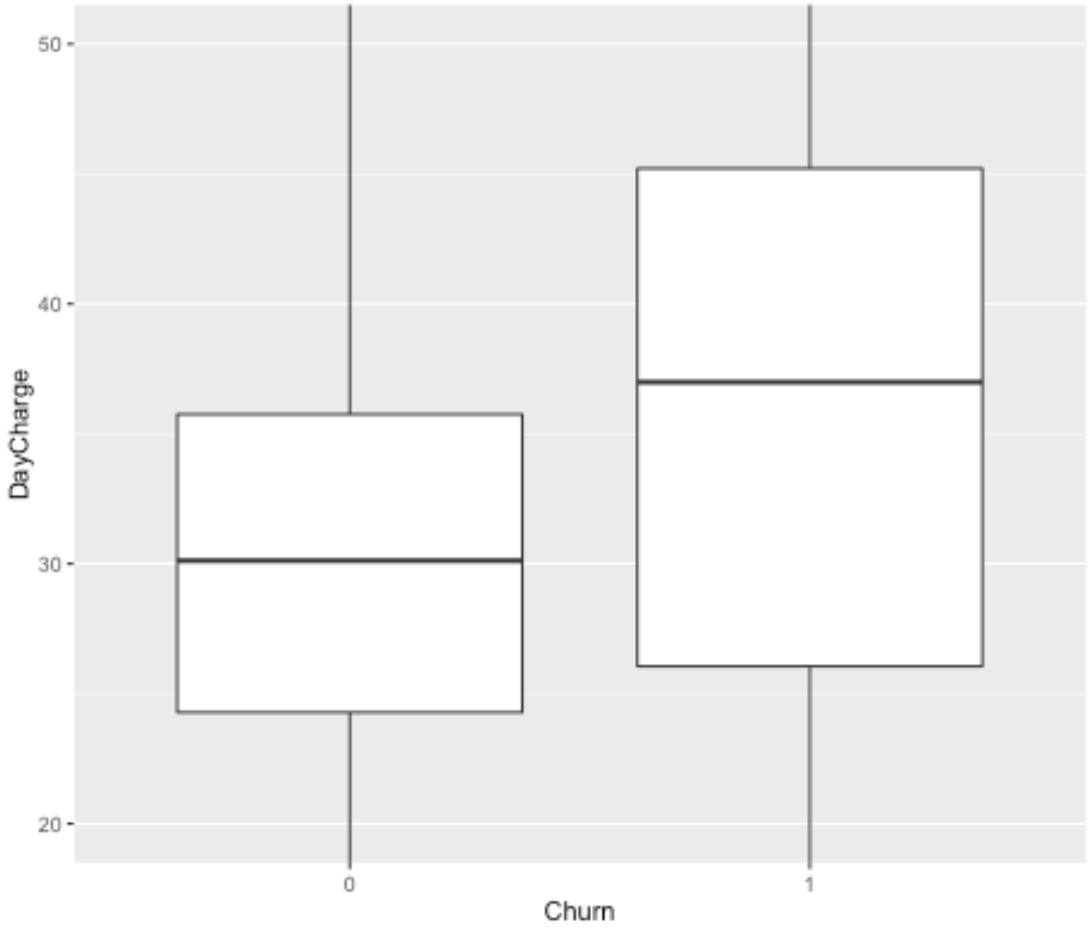
#Churn by Day Charges

#I looked at the stat summary for Day Charges before plotting boxplots so I can set the limits for my Y axis.

by(churn$DayCharge,churn$Churn, summary)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   24.28   30.12   29.78   35.75   53.65
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   26.06   36.99   35.18   45.21   59.64
```

#The boxplots below represent Churn by Day Charge. The plot mirrors what we saw in the Churn by Day Minutes plot – The median for Churners is almost identical to the 3rd quartile for non churners.



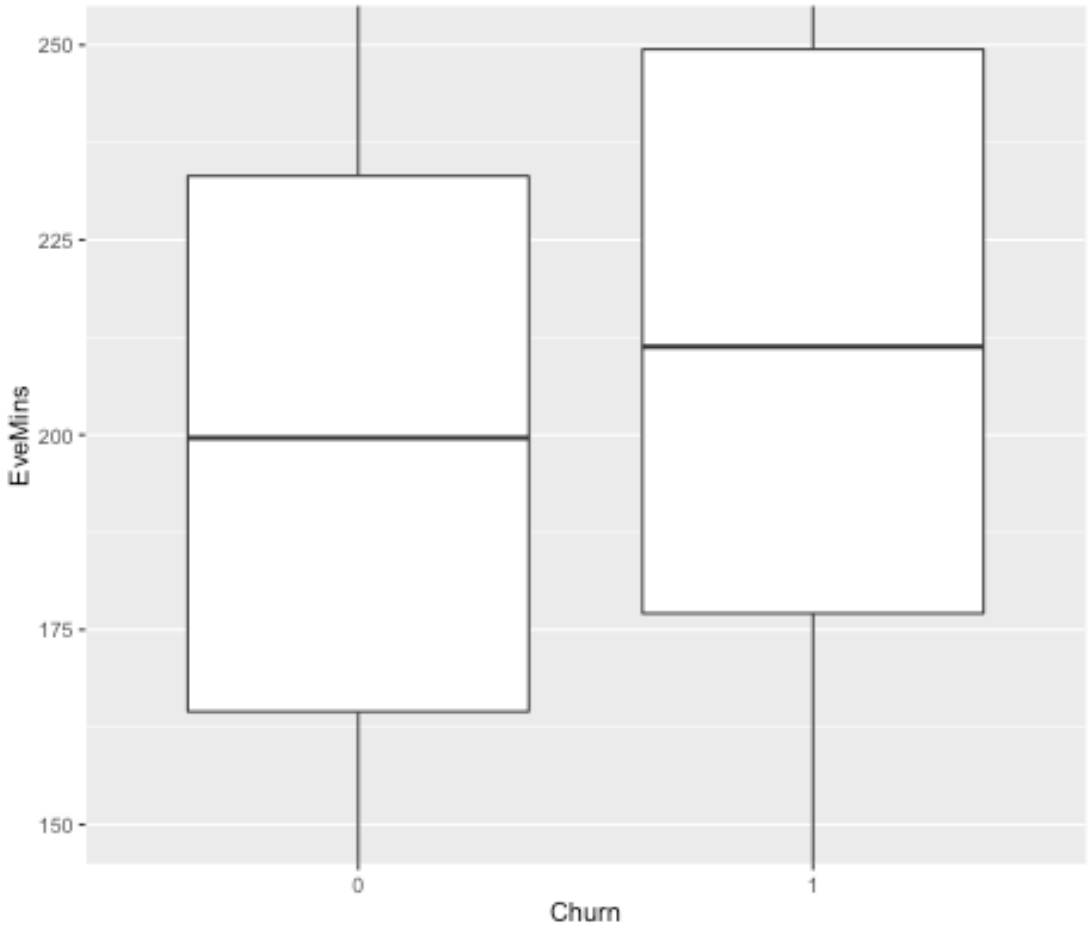
Churn by Day Charges

The higher day charges for Churners isn't really surprising considering we know from the Churn by Day Minutes plot that churners are more frequent users.

#Churn by Evening Minutes

#The boxplot below shows Churn by Evening Minutes. Churners appear to have a higher median and evening minute usage.

```
ggplot(aes(Churn,EveMins),data=subset(churn,!is.na(Churn)))
+geom_boxplot()
+coord_cartesian(ylim=c(150,250))
```



Churn by Evening Minutes

#Lets look at the boxplots closer.

by(churn$EveMins,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   164.5   199.6   199.0   233.2   361.8
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   70.9   177.1   211.3   212.4   249.4   363.7
```

Although there is some difference, I cannot say if this would be enough to influence churn behavior. The higher median for Churners by Evening Minutes is not surprising.

#Churn by Evening Calls

#Lets look at Churn by Evening Calls. The summary is very similar for both Churners and non churners.

by(churn$EveCalls,churn$Churn,summary)

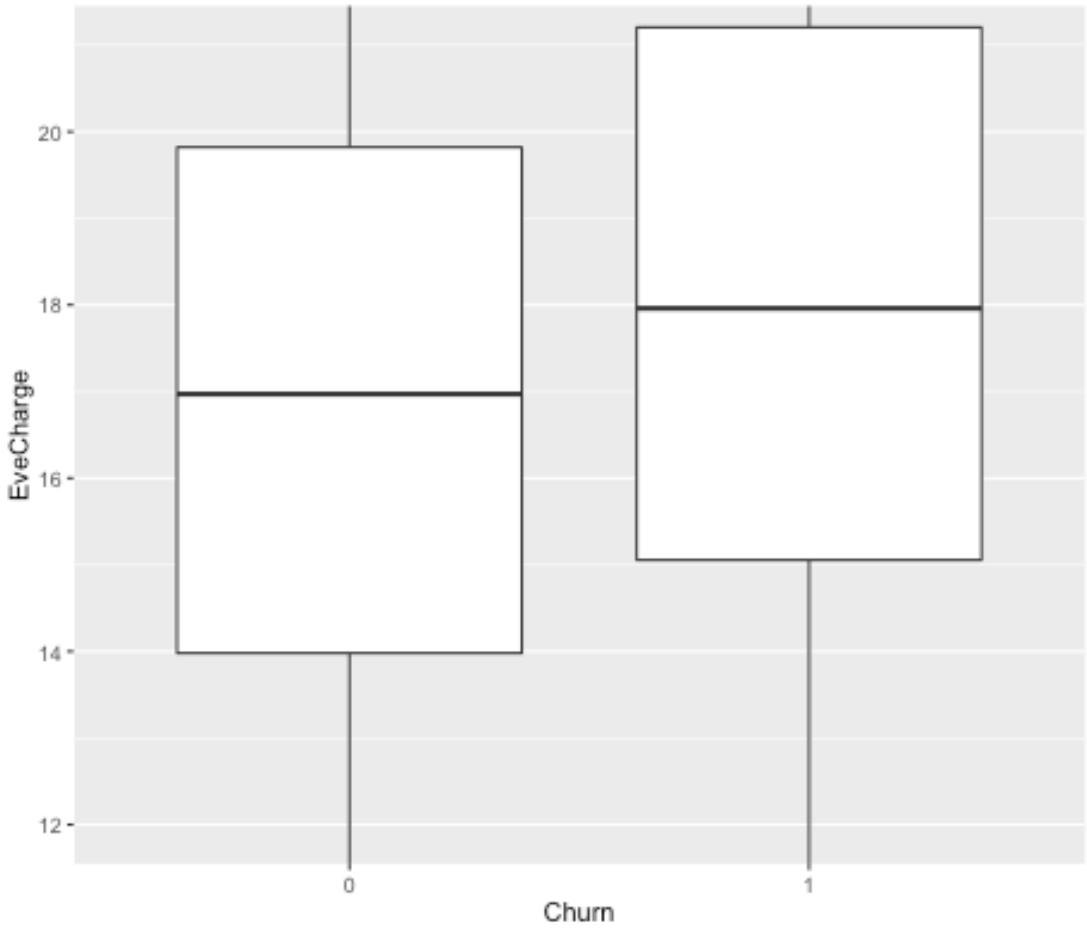
```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0      87     100     100     114     170
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   48.0    87.0   101.0   100.6   114.0   168.0
```

Both Churners and non churners display similar call volumes, but higher evening minutes for churners indicating longer calls for Churners. We can expect the median for Evening Charges to be slightly higher for Churners with these findings.

#Churn by Evening Charges

#The boxplot below shows Churn by Evening Charges. As noted earlier, evening chargers appear to be higher for Churners.

```
ggplot(aes(Churn,EveCharge),data=subset(churn,!is.na(Churn)))
+geom_boxplot()
+coord_cartesian(ylim=c(12,21))
```



Churn by Evening Charges

#A closer look shows a slightly higher median for Churners.

by(churn$EveCharge,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   13.98   16.97   16.92   19.82   30.75
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.03   15.06   17.96   18.05   21.20   30.91
```

#Churn by Night Minutes

#The stat summary below does not show a distinct difference between churners and non churners by night minutes. Night minutes for Churners only appear slightly higher.

by(churn$NightMins,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   23.2   165.9   200.2   200.1   234.9   395.0
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   47.4   171.2   204.8   205.2   239.8   354.9
```

#Churn by Night Calls

#The same goes for Night Calls. The maximum Range for non churners is actually lower than for Churners

by(churn$NightCalls,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   33.0    87.0   100.0   100.1   113.0   175.0
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   49.0    85.0   100.0   100.4   115.0   158.0
```

#Churn by Night Charges

#Night Charges appear identical for both churners and non churners

by(churn$NightCharge,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.040   7.470   9.010   9.006  10.570  17.770
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.130   7.705   9.220   9.236  10.800  15.970
```

The stat summaries for International Minutes, Calls and Charges also did not appear to have distinct differences for churners and non churners.

#Churn by International Minutes

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.40   10.20   10.16   12.00   18.90
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     8.8    10.6    10.7    12.8    20.0
```

#Churn by International Calls

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   4.000   4.533   6.000  19.000
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   4.000   4.164   5.000  20.000
```

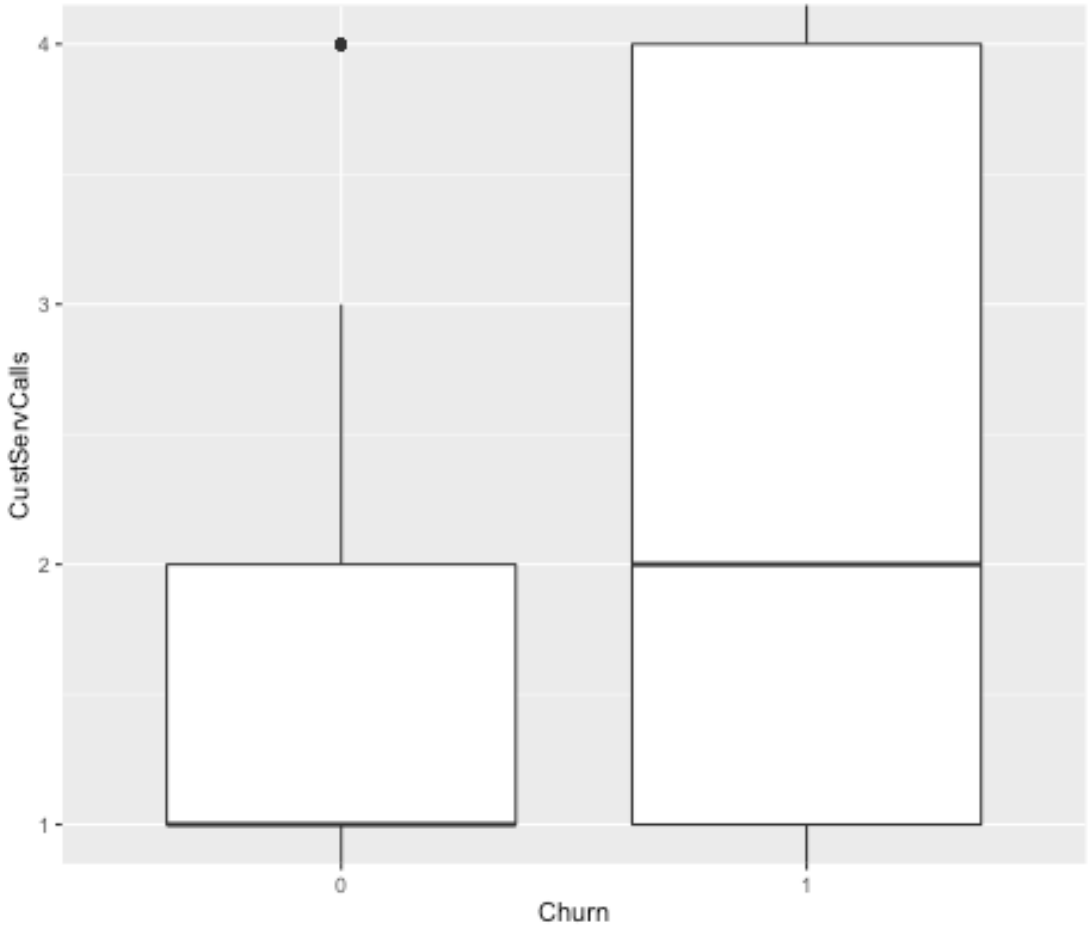
#Churn by International Charges

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.270   2.750   2.743   3.240   5.100
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.54    2.38    2.86    2.89    3.46    5.40
```

#Churn by Customer Service Calls

#The boxplot below shows Churn by Customer Service Calls. The call volumes for Churners are top skewed and have a distinctly higher median.

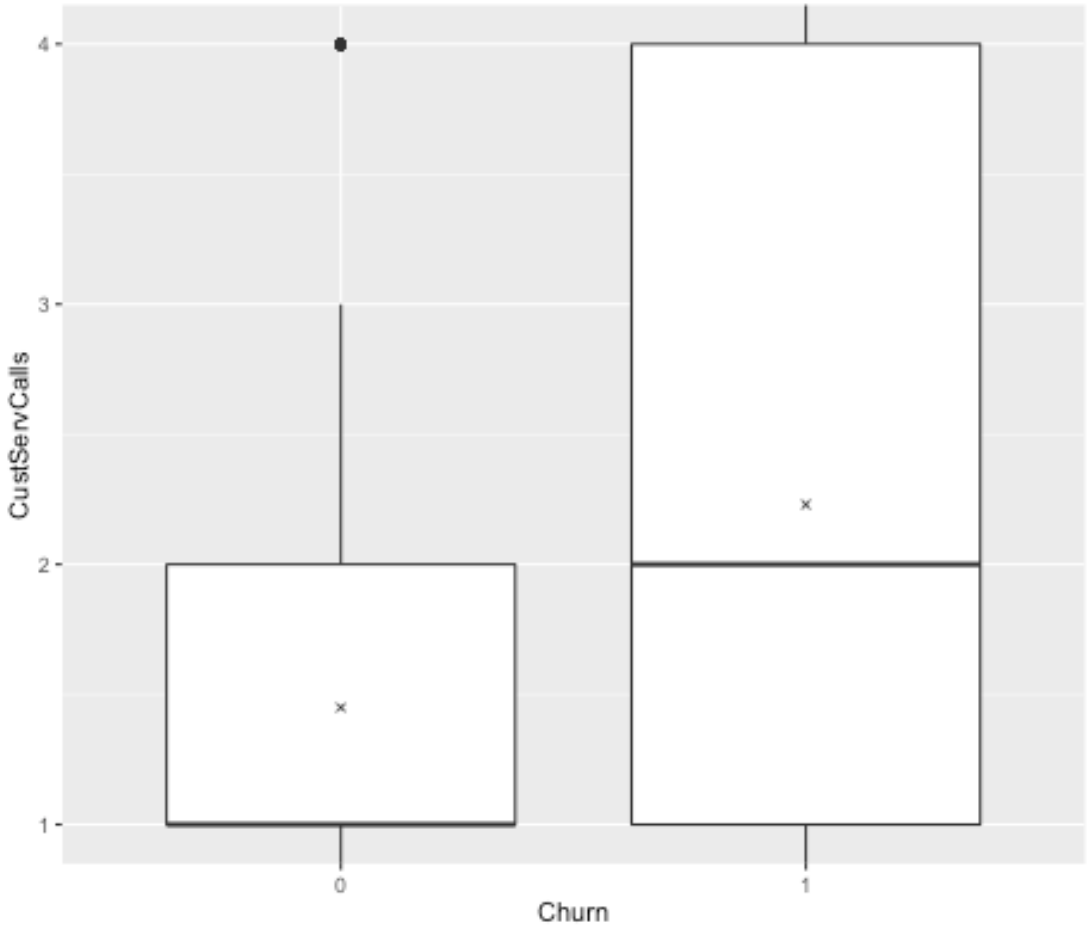
```
ggplot(aes(Churn,CustServCalls),data=subset(churn,!is.na(Churn)))
+geom_boxplot()
+coord_cartesian(ylim=c(1,4))
```



Churn by Customer Service Calls

#Plotting the mean on the boxplots showing top skewed plots

```
ggplot(aes(Churn,CustServCalls),data=subset(churn,!is.na(Churn)))
+geom_boxplot()
+stat_summary(fun.y=mean,geom="point",shape=4)
+coord_cartesian(ylim = c(1,4))
```



Churn by Customer Service Calls

#Lets take a closer look at the numbers represented in the boxplot

```
by(churn$CustServCalls,churn$Churn,summary)
```

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    1.00    1.45    2.00    8.00
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    2.00    2.23    4.00    9.00
```

50% of Churners appear between 1 to 4 customer calls, half of whom made between 2 to 4 calls to Customer Service. Compared to non churners, 75% of whom made 3 or less calls, it seems like Churners made several customer calls before they moved on. This is expected and I would think most calls made by Churners remained largely unresolved

#Churn by Total Calls

#Total Calls is the sum of Morning, Evening and Night Calls. We can see that churners and non churners don't have a distinct difference.

by(churn$totalcalls,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  188.0   277.0   301.0   300.4   323.0   410.0
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  204.0   278.5   302.0   302.3   326.5   407.0
```

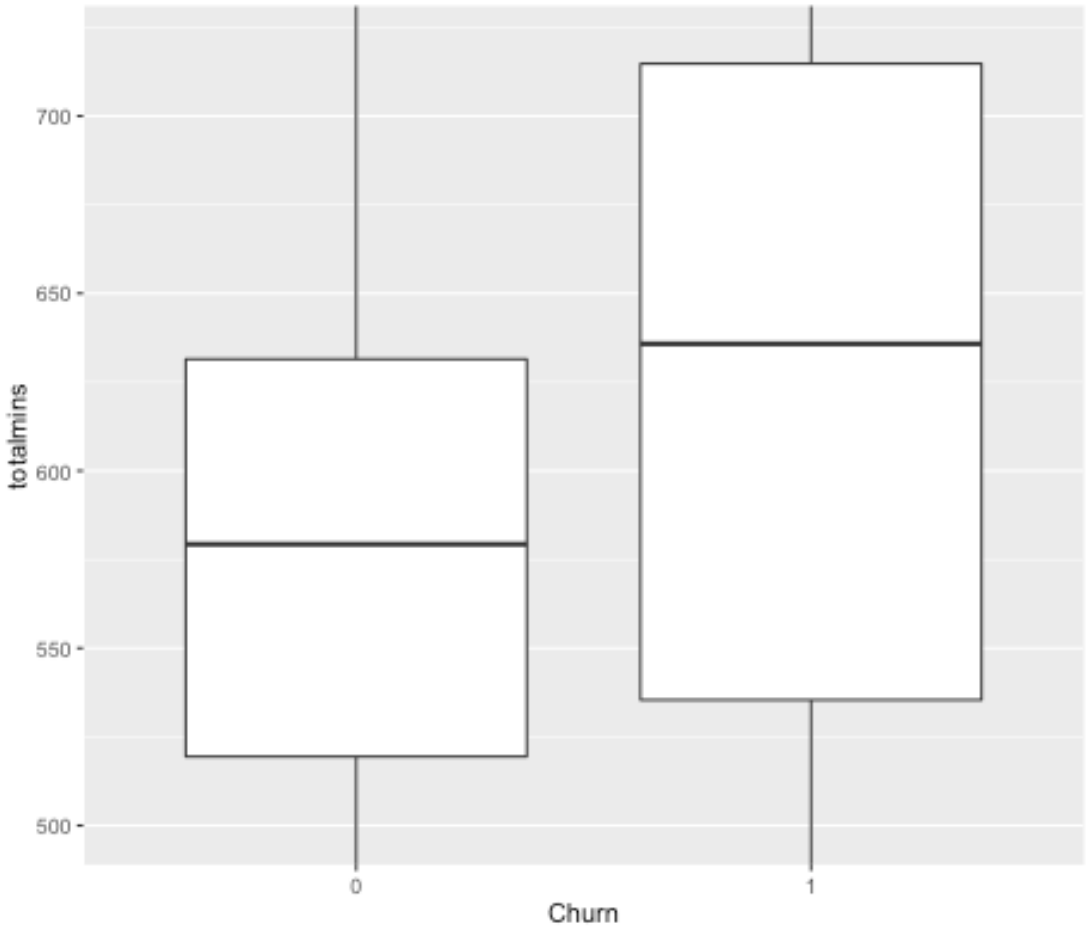
#Churn by Total Minutes

#Total Calls is the sum of Morning, Evening and Night minutes. Since it is a factor of day minutes, it is not surprising to see a larger median for Churners.

by(churn$totalcalls,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  275.2   519.4   579.3   574.4   631.5   831.0
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  311.7   535.4   635.8   624.6   714.8   876.9
```

#The boxplots below represent Churn by total minutes and as expected, have a higher median. Churners are also more widely distributed than non churners.

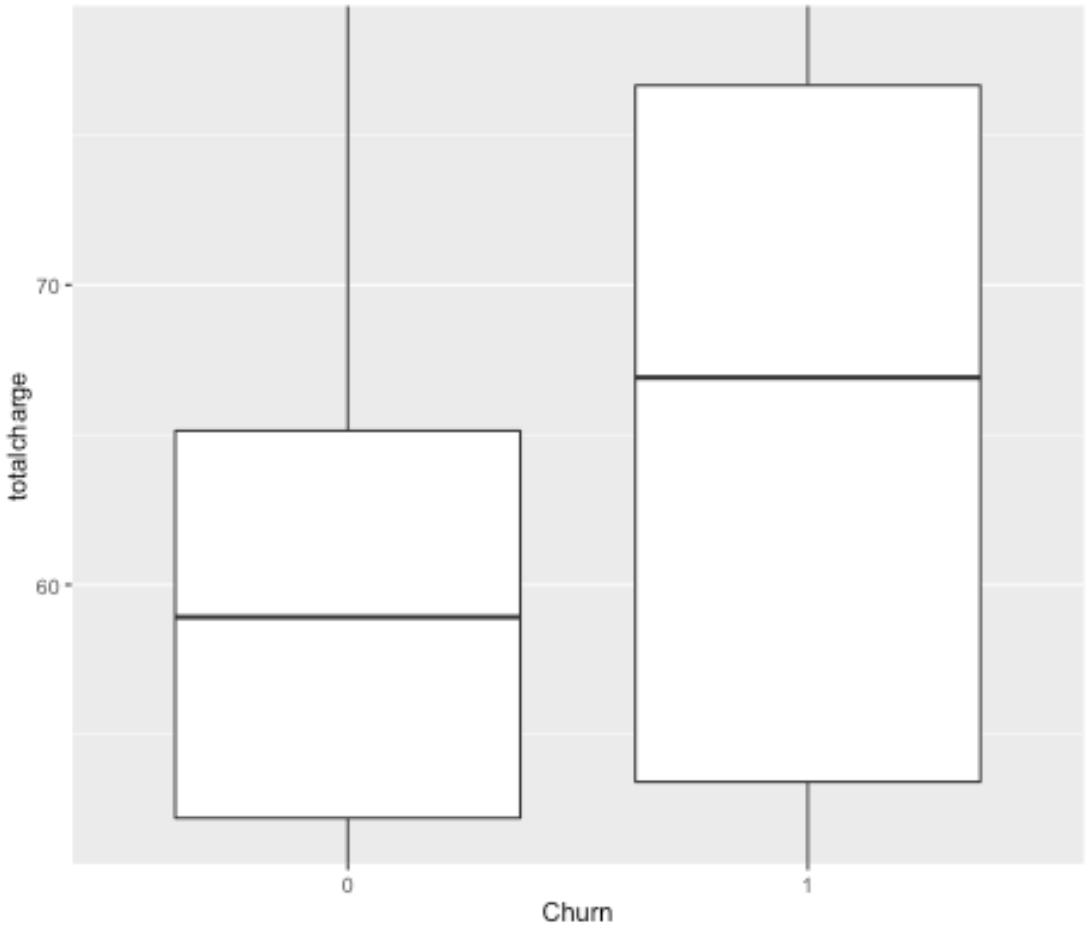


Churn by Total Minutes

#Churn by Total Charges

#Total Charges is the sum of Morning, Evening, Night and International charges. We expect to see higher values in the boxplots for churners because day charges displayed the same behavior.

```
ggplot(aes(Churn,totalcharge),data=subset(churn,!is.na(Churn)))
+geom_boxplot()
+coord_cartesian(ylim=c(52,78))
```



Churn by Total Charges

#Lets look at the summary. The median for Churners by Total Charges is almost at the 3rd quartile for non churners.

by(churn$totalcharge,churn$Churn,summary)

```
churn$Churn: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.25   52.22   58.92   58.45   65.14   87.29
------------------------------------------------
churn$Churn: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.93   53.42   66.91   65.36   76.66   96.15
```

Logistic Regression Modelling

We'll now look at the process of building a model for our dataset. We'll divide our dataset into train and test sets. The train set is used to build our model and the test set will be used to test our model on.

```
#Splitting data into Test and Train Sets.
#We'll retain 60% of the dataset for the train set.
library(caTools)
split=sample.split(churn$Churn,SplitRatio = 0.60)
train<-subset(churn, split==TRUE)
test<-subset(churn, split==FALSE)

#Logistic Regression
#We'll use the glm function with Churn as our response variable on the training dataset with the binomial family. The following were used as predictor variables after I tried a series of models.
churnLog1 <-  glm (Churn~IntlPlan+ VMailPlan+ VMailMessage+ IntlCalls+ CustServCalls+ totalcharge, data= train, family = binomial)
summary(churnLog1)
```

```
Call:
glm(formula = Churn ~ IntlPlan + VMailPlan + VMailMessage + IntlCalls +
    CustServCalls + totalcharge, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9822  -0.5318  -0.3469  -0.1972   3.3368

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.019618   0.530026 -13.244  < 2e-16 ***
IntlPlan1      2.056477   0.190077  10.819  < 2e-16 ***
VMailPlan1    -2.537319   0.769206  -3.299 0.000972 ***
VMailMessage   0.047925   0.023490   2.040 0.041325 *
IntlCalls     -0.104066   0.032587  -3.193 0.001406 **
CustServCalls  0.506157   0.049823  10.159  < 2e-16 ***
totalcharge    0.075917   0.007619   9.964  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1655.7  on 1999  degrees of freedom
Residual deviance: 1310.2  on 1993  degrees of freedom
AIC: 1324.2

Number of Fisher Scoring iterations: 6
```

Notice that Voice Mail Plans and International Calls have negative values indicating an inverse relationship with Churn. It is possible the service has a good international plan and so customers using international minutes, are resistant to move away. Signing up for the Voice Mail Plan may have add ons that customers do not want to lose by moving away.

#We'll now use the predict function on the test data, and compare the models predictions with the actual churn results from the test data using a confusion matrix.

```
predictTest<-predict(churnLog1, type="response",newdata=test)
table(test$Churn,predictTest>0.5)
```

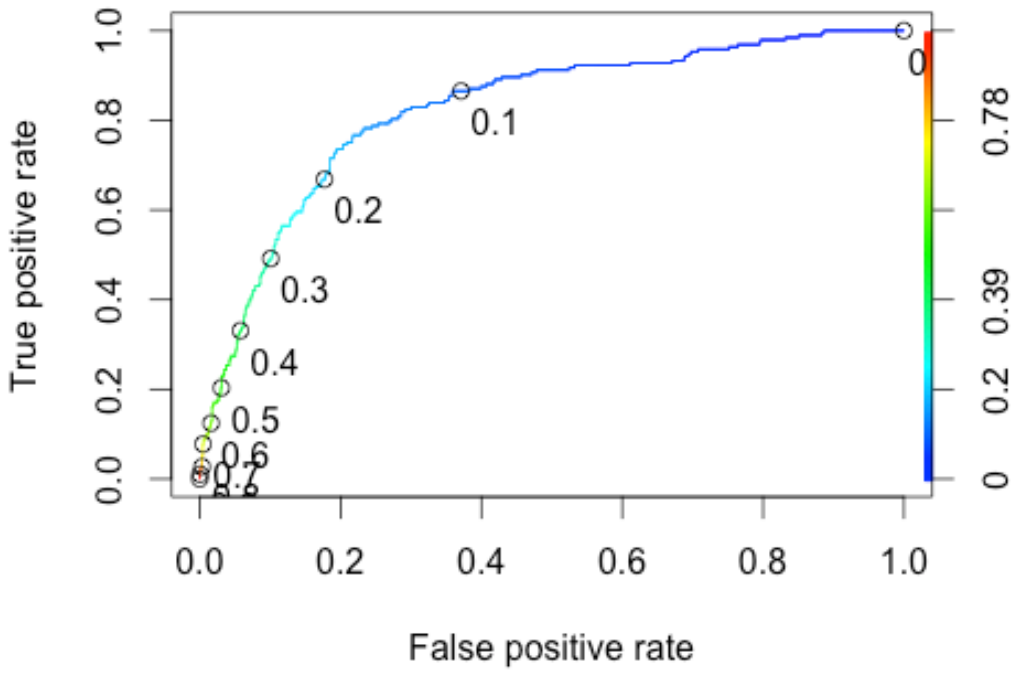
```
   FALSE  TRUE
0   1105    35
1    154    39
```

#We can tell the accuracy of the model to be 0.8582146 from the above table.

#Receiver Operating Characterisitc (ROC) Curve

#We will now use the ROC Curve to describe sensitivity/specificity and the Area under Curve (AUC) on the testing set

```
library(ROCR)
ROCRtestpred<-prediction(predictTest,test$Churn)
ROCRtestperf<-performance(ROCRtestpred,"tpr","fpr")
plot(ROCRtestperf,colorize=TRUE,print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
as.numeric(performance(ROCRtestpred, "auc")@y.values)
```



ROC Curve

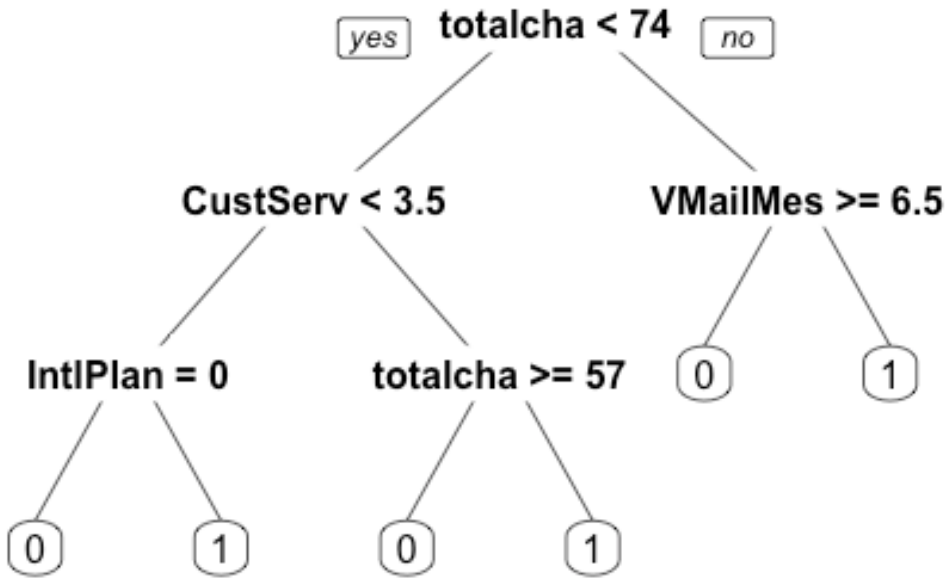
#The value for AUC is 0.8239433. The ROC Curve shows the tradeoff between sensitivity on the Y axis and specificity on X axis.

#Decision Trees

#We'll now look at plotting Decision Trees.

```
library(rpart)
library(rpart.plot)

ChurnTree<- rpart(Churn~IntlPlan+ VMailPlan+ VMailMessage+ IntlCalls+
CustServCalls+ totalcharge,data=train,method="class",minbucket=30,parms =
list(prior=c(0.5,0.5)))
prp(ChurnTree)
```



Decision Tree Plot

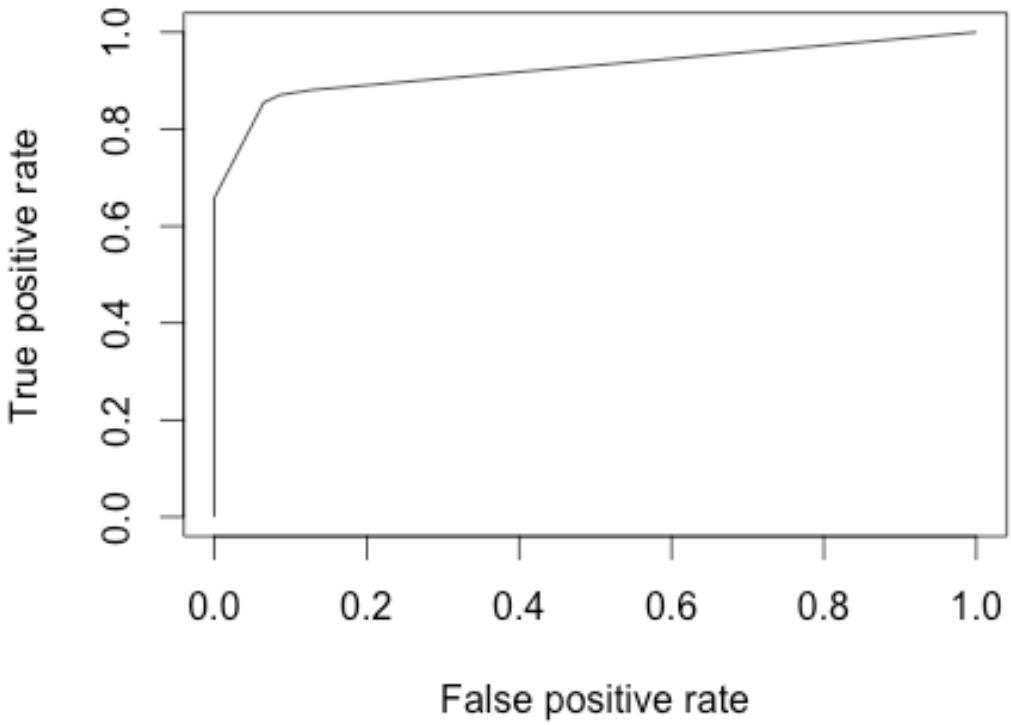
#Confusion Matrix for Decision Tree. This gives us an accuracy of 0.9234809

```
predictCART<-predict(ChurnTree,newdata=test,type="class")
table(test$Churn,predictCAR
```

```
  predictCART
     0    1
0 1066   74
1   28  165
```

#ROC Curve

```
predictROC<-predict(ChurnTree,newdata=test)
pred<-prediction(predictROC[,2],test$Churn)
perf<-performance(pred,"tpr","fpr")
plot(perf)
```



ROC Plot for Decision Tree

#The AUC value is 0.9241

```
as.numeric(performance(pred, "auc")@y.values)
```

Conclusions.
I would like to explore the following for further research.

- Customers with International Calls showed decreased rates of Customer Churn. As mentioned before, the service may have an International Plan that is better than other options in the market. This could be cheap rates or better connectivity, etc.

- However, customers with International Plans increased churn rate. This is possibly due to a base retainer fee that is billed to avail International Services irrespective of actual usage.

- It seems that this could happen if these customers were not frequent travelers or the plan had a minimum lock in period that was excessive for these customers.

- These could also be attributed to users who only signed up for the service because they were travelling on occasion and deactivated the service on return.

- We have previously seen customers with international minutes show decreased churn rates, and this could be an indicator of the service provider's popularity when traveling.

- The Voice Mail Plan also has an inverse relationship with customer Churn which may be attributed to an enhanced plan with the Voice Mail sign up. For example, added minutes or a family plan on the Voice Mail Plan would make it less easy to sign away.

- Customers with Voice Mail Messages however, displayed higher churn rates. It is likely that the actual voice mail system is not up to standard because of dropped messages, wrong messages, inability to edit or access messages, etc.

- Customer Service Calls increased churn rates. This seems like an acceptable find. Increased customer service calls likely indicate dissatisfaction with the service leading to churn.

- Customer Churn also increased with Total Charges. I don't think this is restricted to this particular telecom provider, and could be an industry trend. It seems like this could be a customer going over budget on their plans or being billed for services they did not accept, and decided to move.

- Comparing AUC Values (0.824 in Logistic Regression vs 0.924 in Decision Tree), the Decision Tree Model performs better and is selected for further prediction.

Further Research

Although we have uncovered some relationships, a more detailed analysis will required more information.
Further research should include the predictor variables, and timelines, calendar dates, customer demographics, etc.

Recommendations

There is a disconnect between signing up for an International Plan, and using International Minutes. The company should look at the data for International Plans more closely and ascertain why they have increased churn rates. I propose doing away with any lock in periods and converting any leftover international minutes to local minutes to reduce churn amongst occasional travelers.

I recommend a deeper analysis of Customer Service Calls. Increased calls are leading to churn, and finding out why is an important step towards retention. The company should first focus on the most common problems and slowly tackle a wider range. Customer Service should also be toll free and any calls that are left unresolved after defined timelines should be escalated to bring the median for churners down.

Churn by Total Charges is a trickier metric to fix, and is likely an industry issue. I recommend communicating warnings when a customer exceeds their minutes or unfamiliar activity is detected on their usage. However, this may have limited impact and a spillover of pending charges to other telecom providers will be a larger deterrent. The nature of competition in this industry may not allow that though and something akin to a credit score may be the middle path.

The company should look at the Voice Mail System and if needed, commission a redesign of the system. The Voice Mail is accepted as a regular feature and not really a benefit, and so customers may take this for granted only to be disappointed later expressed in their churn.

It seems like International Calls have an inverse relationship with Churn. This could indicate a strong position in terms of International service with its customers. I recommend tailoring ad campaigns to immigrants and the international community.