Regression Analysis

Regression Analysis in Practice

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Customer Churn Analysis in the Telecom Sector

Georgia Tech

About This Lesson

Georgia Tech

Customer Churn Analysis

Customer Churn is of great interest in industries where revenues are heavily dependent on subscriptions.

Dataset: Customer data for 7,043 telecom clients, all located in CA, USA.

Data Source: IBM Business Analytics Community

Acknowledgement: *This example was prepared with support from students in the Masters of Analytics program, including Jared Babcock, Rishi Bubna, Marta Bras, Aymee Garcia Lopez Gavilan, and Artur Bessa Cabral.*

Georgia Tech

Response & Predicting Variables

Response variables:

- *Churn Value:* 1 = the customer left the company. 0 = the customer remained with the company.

Predicting variables:

- **Demographics:** 4 variables including customer's gender (*Gender*), marital status (*Partner*) among others.
- **Location:** 7 variables including customer's primary residence ZIP Code (*Zip Code*), latitude (*Latitude*) among others.
- **Services:** 15 variables including customer's subscriptions to home phone (*Phone Services*), internet (*Internet services*), tech support (*Tech Support*) among others services.
- **Status:** 6 variables including customer's ID (*CustomerID*), reason for leaving the company (*Churn Reason*), customer's lifetime value (CLTV) among others.

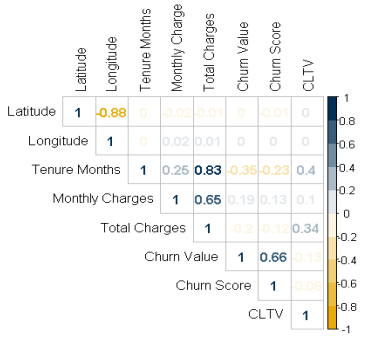
Georgia Tech

Objective and Methods

- Predict which customers are likely to churn.
  - Logistic Regression
  - K Nearest Neighbors
  - Decision Tree
  - Random Forest

Exploratory Data Analysis in R

```
## Correlation among the numeric variables
# Select numerical variables
dat.num <- na.omit(dat[ , which(sapply(dat, is.numeric))])

# Create correlation matrix
corr <- cor(dat.num)

# Create correlation plot
col <- colorRampPalette(c(buzzgold,"white", gtblue))(10)
corrplot(corr, method = "number", type = "upper",
         tl.col="black", col = col)
```



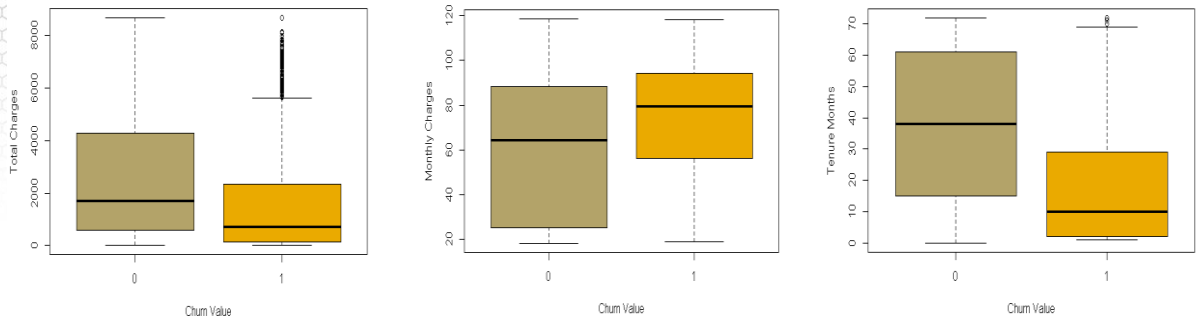
There appears to be strong correlation among some of the predicting variables.

Exploratory Data Analysis in R (cont'd)

```
## Relationship between binary response and numerical variables

par(mfrow=c(2,2))
boxplot(Total.Charges ~ Churn.Value, main="", xlab="Churn Value", ylab="Total Charges",
col=c(techgold,buzzgold), data=dat)

boxplot(Monthly.Charges ~ Churn.Value, main="", xlab="Churn Value", ylab="Monthly Charges",
col=c(techgold,buzzgold), data=dat)

boxplot(Tenure.Months ~ Churn.Value, main="", xlab="Churn Value", ylab="Tenure Months",
col=c(techgold,buzzgold), data=dat)
```

Georgia Tech

Exploratory Data Analysis in R (cont'd)

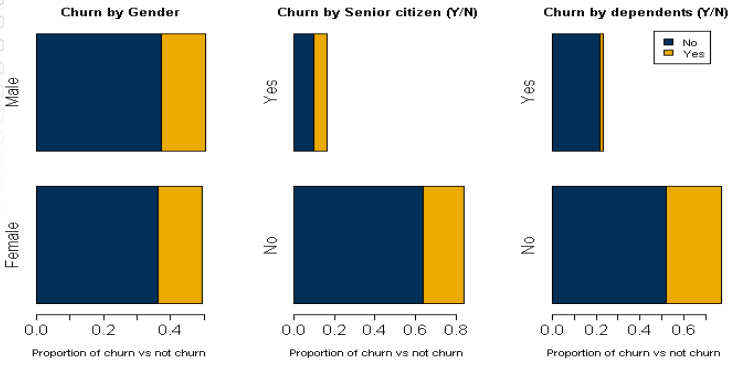


Customers that remain with the company appear to have higher total charges and tenure months but lower monthly charges than customers that have churned.

Georgia Tech

Exploratory Data Analysis in R (cont'd)

```
## Relationship between binary response and categorical variables
par(mfrow=c(1,3))

tb_obgender = xtabs(~dat$Churn.Value+ dat$Gender)
barplot(prop.table(tb_obgender),axes=T,space=0.3,  cex.axis=1.5, cex.names=1.5,
        xlab="Proportion of churn vs not churn",
        horiz=T, col=c(gtblue,buzzgold),main="Churn by Gender")

tb_citizen = xtabs(~dat$Churn.Value+ dat$Senior.Citizen)
barplot(prop.table(tb_citizen),axes=T,space=0.3, cex.axis=1.5, cex.names=1.5,
        xlab="Proportion of churn vs not churn",
        horiz=T, col=c(gtblue,buzzgold),main="Churn by Senior citizen (Y/N)")

tb_Dependents = xtabs(~dat$Churn.Value+ dat$Dependents)
barplot(prop.table(tb_Dependents),axes=T,space=0.3,cex.axis=1.5, cex.names=1.5,
        xlab="Proportion of churn vs not churn ",
       horiz=T, col=c(gtblue,buzzgold),main="Churn by dependents (Y/N)",
       legend.text = c("No", "Yes"))
```



There seems to exist significant differences in the proportions for each group in the predicting variables *Senior Citizen* and *Dependents*.

Summary