Telcommunications Churn Analysis

Written by Savahnna L. Cunningham

Introduction

You are an analyst for a telecommunications company that is concerned about the number of customers leaving their landline business for cable competitors. The company needs to know which customers are leaving and attempt to mitigate continued customer loss. You have been asked to analyze customer data to identify why customers are leaving and potential indicators to explain why those customers are leaving so the company can make an informed plan to mitigate further loss.

Methods & Analysis

Data Gathering

```
# Load the Data
churn <- read_csv("~/Desktop/MSDA Portfolio/4.Data Mining/Data/2. Processed Data/Telco_Churn.csv")
```

Summary table representing the structure of the dataset.The Telco_Churn dataset contains 7043 rows (customers) and 21 variables (features). The Churn feature is the target variable and is of a discrete categorical data type based on if a customer has left the company or not.

```
str(churn)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ Gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : num  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
```

```
##  - attr(*, "spec")=
##   .. cols(
##   ..   customerID = col_character(),
##   ..   Gender = col_character(),
##   ..   SeniorCitizen = col_double(),
##   ..   Partner = col_character(),
##   ..   Dependents = col_character(),
##   ..   tenure = col_double(),
##   ..   PhoneService = col_character(),
##   ..   MultipleLines = col_character(),
##   ..   InternetService = col_character(),
##   ..   OnlineSecurity = col_character(),
##   ..   OnlineBackup = col_character(),
##   ..   DeviceProtection = col_character(),
##   ..   TechSupport = col_character(),
##   ..   StreamingTV = col_character(),
##   ..   StreamingMovies = col_character(),
##   ..   Contract = col_character(),
##   ..   PaperlessBilling = col_character(),
##   ..   PaymentMethod = col_character(),
##   ..   MonthlyCharges = col_double(),
##   ..   TotalCharges = col_double(),
##   ..   Churn = col_character()
##   .. )
```

Address Quality & Tidyness Issues

Identify & Remove Missing Data

```
#Id rows
sapply(churn, function(x) sum(is.na(x)))
```

```
##       customerID           Gender    SeniorCitizen          Partner
##                0                0                0                0
##       Dependents           tenure     PhoneService    MultipleLines
##                0                0                0                0
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##                0                0                0                0
##      TechSupport      StreamingTV  StreamingMovies         Contract
##                0                0                0                0
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##                0                0                0               11
##            Churn
##                0
```

Found 11 missing values in the "Total Charges" column. Due to the large size of the dataset, removal of the rows will not have an adverse effect on the prediction models.

```
#remove the rows with missing values
churn <- churn[complete.cases(churn), ]
```

Tidyness Issue: Recode string values

```
# "No Internet Service" --> "No" for 6 variables [11&13].
churn$OnlineBackup <- as.factor(mapvalues(churn$OnlineBackup,
                                          from=c("No internet service"),
                                          to=c("No")))
churn$OnlineSecurity <- as.factor(mapvalues(churn$OnlineSecurity,
                                          from=c("No internet service"),
                                          to=c("No")))
churn$DeviceProtection <- as.factor(mapvalues(churn$DeviceProtection,
                                          from=c("No internet service"),
                                          to=c("No")))
churn$TechSupport <- as.factor(mapvalues(churn$TechSupport,
                                          from=c("No internet service"),
                                          to=c("No")))
churn$StreamingTV <- as.factor(mapvalues(churn$StreamingTV,
                                          from=c("No internet service"),
                                          to=c("No")))
churn$StreamingMovies <- as.factor(mapvalues(churn$StreamingMovies,
                                          from=c("No internet service"),
                                          to=c("No")))

#"No Phone Service" --> "No" for the MultipleLines variable [11&13].
churn$MultipleLines <- as.factor(mapvalues(churn$MultipleLines,
                                          from=c("No phone service"),
                                          to=c("No")))
#(Rokicki, 2012; R Documentation: Levels Attributes)
```

Tidyness Issue: Recode Senior Citizen variable to binary numeric values

```
# Change the values in the SeniorCitizen column from 0 or 1 to "No" or "Yes"
churn$SeniorCitizen <- as.factor(mapvalues(churn$SeniorCitizen,
                                          from=c("0","1"),
                                          to=c("No","Yes")))
#(Rokicki, 2012; R Documentation: Levels Attributes)
```

Tidyness Issue: Divide Tenure variable into bins

Divide the Tenure variable into bins based on the range of the numeric data. Change the numeric type of the Tenure variable to a factor and add it to the data frame as a new variable tenure_grp. Additionally, five bins will be created representing the number of months a customer has been with the company.

```
#variable range?
min(churn$tenure)
```

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

```
max(churn$tenure)
```

```
## [1] 72
```

```r
#Bin Creation
#Change numeric variable to a factor.
#Save to data frame as a new variable tenure_grp

churn$tenure_grp <- factor(churn$tenure)

churn$tenure_grp <- ifelse(churn$tenure>=0 & churn$tenure<=12,
                           "0-12", churn$tenure_grp)
churn$tenure_grp <- ifelse(churn$tenure>12 & churn$tenure<=24,
                           "12-24", churn$tenure_grp)
churn$tenure_grp <- ifelse(churn$tenure>24 & churn$tenure<=48,
                           "24-48", churn$tenure_grp)
churn$tenure_grp <- ifelse(churn$tenure>48 & churn$tenure<=60,
                           "48-60", churn$tenure_grp)
churn$tenure_grp <- ifelse(churn$tenure>60,
                           ">60", churn$tenure_grp)
churn$tenure_grp <- factor(churn$tenure_grp,
                           levels = c("0-12",
                                      "12-24",
                                      "24-48",
                                      "48-60",
                                      ">60"))
#(Rokicki, 2012; R Documentation: Levels Attributes)
```

```r
#Drop columns no longer needed for analysis
churn$customerID <- NULL
churn$tenure <- NULL

#rearrange columns so the target var is last
churn<- churn[c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20,19)]
```

Exploratory Data Analysis

Descriptive Statistics

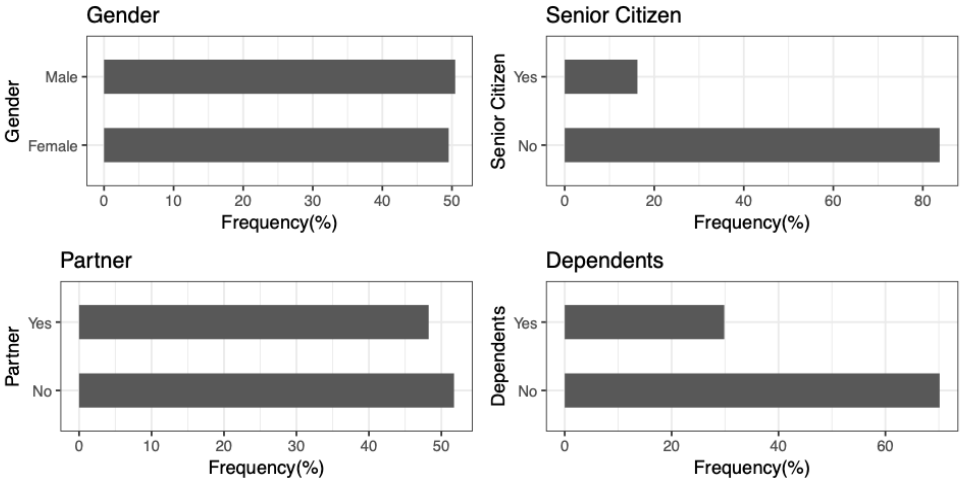
```r
#Statistics that describe the basic features of the dataset.
summary(churn)
```

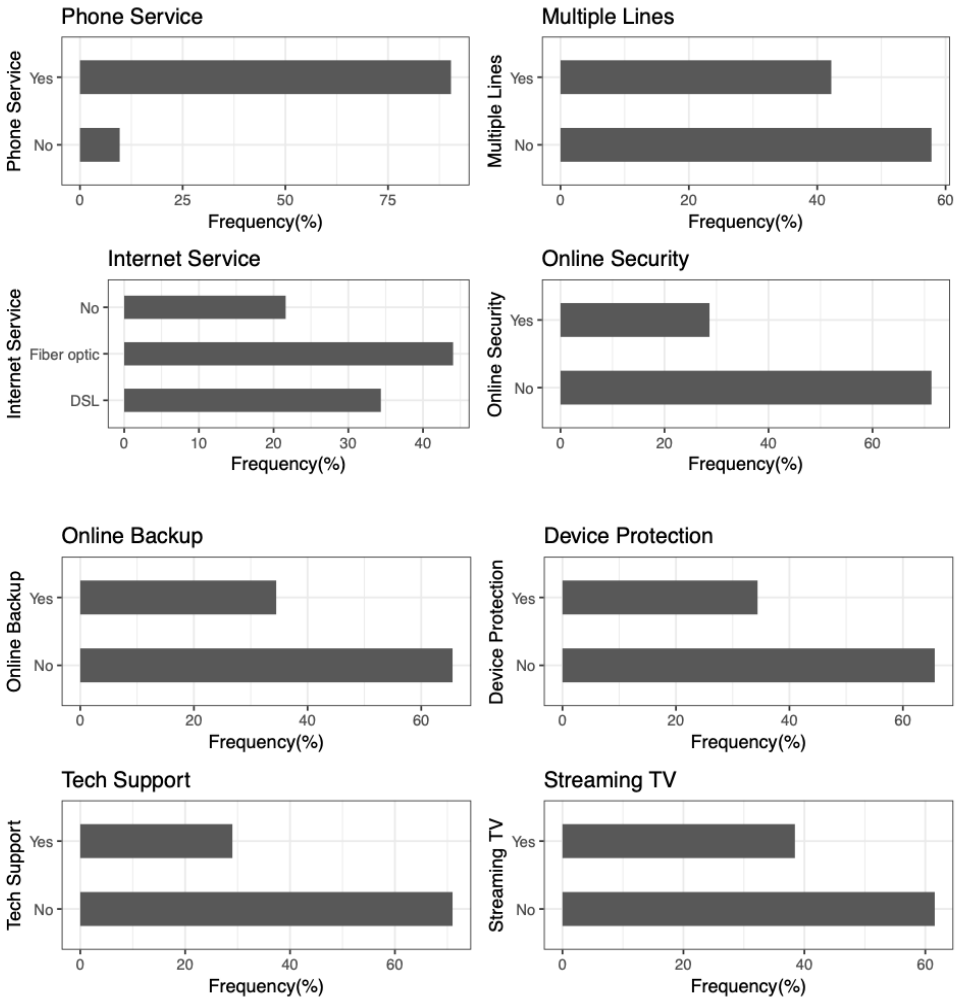
```
##     Gender          SeniorCitizen   Partner           Dependents
##  Length:7032        No :5890      Length:7032        Length:7032
##  Class :character   Yes:1142      Class :character   Class :character
##  Mode  :character                 Mode  :character   Mode  :character
##
##
##
##  PhoneService       MultipleLines InternetService    OnlineSecurity
##  Length:7032        No :4065      Length:7032        No :5017
##  Class :character   Yes:2967      Class :character   Yes:2015
##  Mode  :character                 Mode  :character
##
```

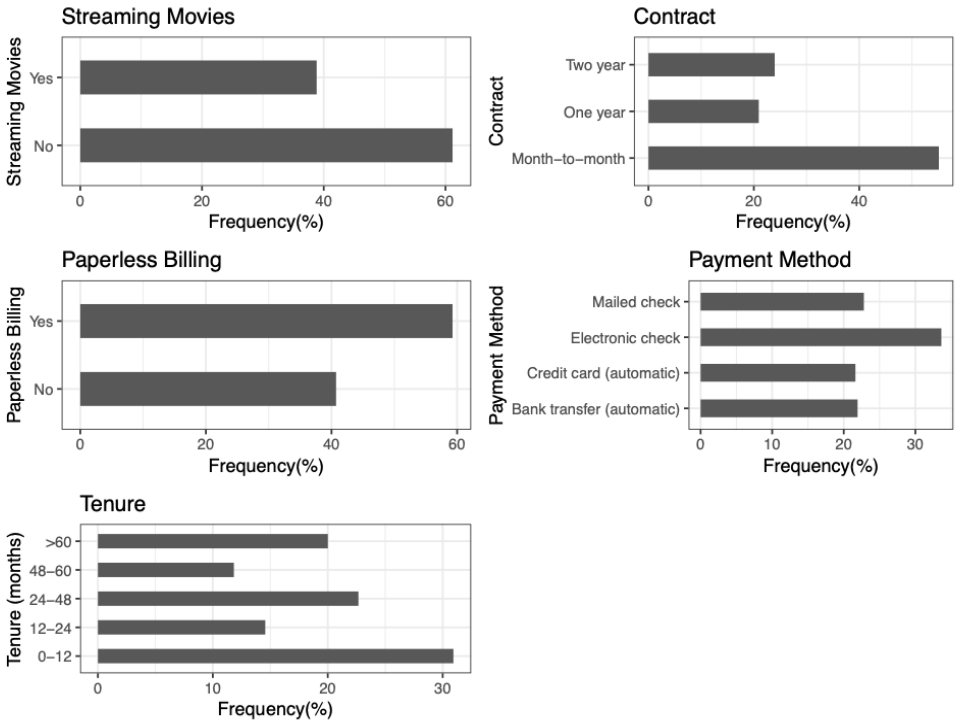
```
##
##
##  OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
##  No :4607     No :4614         No :4992    No :4329    No :4301
##  Yes:2425     Yes:2418         Yes:2040    Yes:2703    Yes:2731
##
##
##
##
##    Contract         PaperlessBilling   PaymentMethod      MonthlyCharges
##  Length:7032        Length:7032        Length:7032        Min.   : 18.25
##  Class :character   Class :character   Class :character   1st Qu.: 35.59
##  Mode  :character   Mode  :character   Mode  :character   Median : 70.35
##                                                           Mean   : 64.80
##                                                           3rd Qu.: 89.86
##                                                           Max.   :118.75
##   TotalCharges    tenure_grp      Churn
##  Min.   :  18.8   0-12 :2175   Length:7032
##  1st Qu.: 401.4   12-24:1024   Class :character
##  Median :1397.5   24-48:1594   Mode  :character
##  Mean   :2283.3   48-60: 832
##  3rd Qu.:3794.7   >60  :1407
##  Max.   :8684.8
```

Univariate Analysis

Categorical Variable Bar Graphs



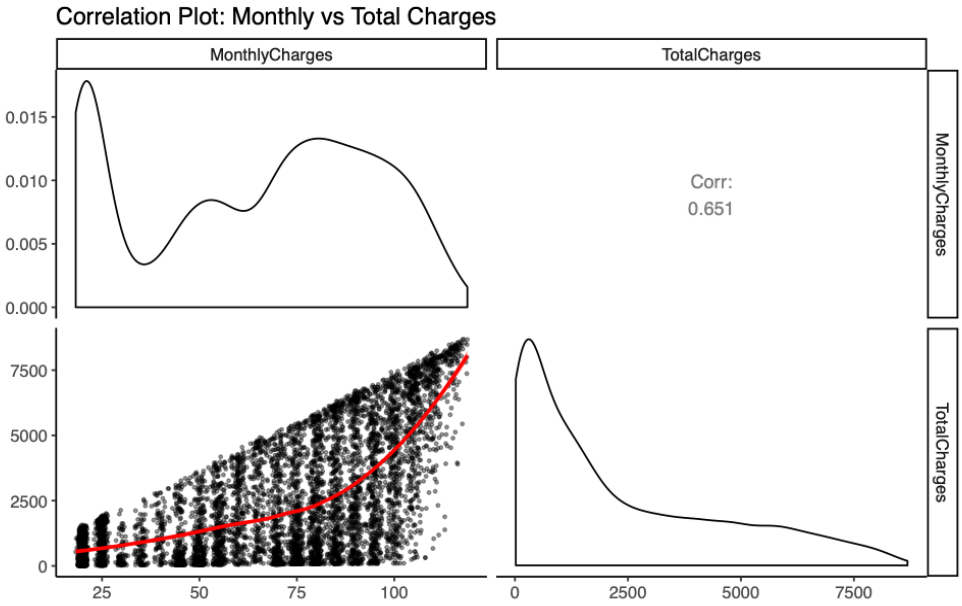

Phone Service
Phone Service
Yes
No
0
25
50
75
Frequency(%)
Multiple Lines
Multiple Lines
Yes
No
0
20
40
60
Frequency(%)
Internet Service
Internet Service
No
Fiber optic
DSL
0
10
20
30
40
Frequency(%)
Online Security
Online Security
Yes
No
0
20
40
60
Frequency(%)
Online Backup
Online Backup
Yes
No
0
20
40
60
Frequency(%)
Device Protection
Device Protection
Yes
No
0
20
40
60
Frequency(%)
Tech Support
Tech Support
Yes
No
0
20
40
60
Frequency(%)
Streaming TV
Streaming TV
Yes
No
0
20
40
60
Frequency(%)



Summary

The Telco_Churn dataset contains 7032 customer samples comprised of 19 independent variables that affect a customer's retention. There are 16 categorical and 2 quantitative variables contained in the dataset. Bar charts were used to better understand the frequency(count) percentage of the independent variables. The 19 independent variables are the primary features of interest and are critical to understanding how they correlate to customer churn status.

Bivariate Analysis



Summary

A Pearson correlation coefficient was computed to assess the relationship between the Monthly and Total charges customers were billed. There was a moderate, positive correlation between the two variables, r = 0.65, n = 7,032. Increases in monthly charges were correlated with increases in total charges. A correlation plot summarizes the results. Due to the positive correlation the Total Charges variable will be removed from the dataset to help model accuracy.

The categorical variables have a wide range in count distribution. Therefore, no variable will be discarded, and all will contribute to the machine learning algorithms.

Mathematical Model Preprocessing

```
#Copy the dataset
churn_clean <- churn
```

Encoding Categorical Variables

Machine learning algorithms require numerical input and output variables.

```
churn_clean$tenure_grp = factor(churn_clean$tenure_grp,
                        levels = c('0-12', '12-24', '24-48','48-60','>60'),
```

```
                          labels = c(1, 2, 3, 4, 5))
churn_clean$PaymentMethod = factor(churn_clean$PaymentMethod,
                          levels = c('Mailed check','Electronic check',
                                     'Credit card (automatic)',
                                     'Bank transfer (automatic)'),
                          labels = c(1, 2, 3, 4))
churn_clean$Contract = factor(churn_clean$Contract,
                          levels = c('One year','Two year','Month-to-month'),
                          labels = c(1, 2, 3))
churn_clean$InternetService = factor(churn_clean$InternetService,
                          levels = c('No','DSL','Fiber optic'),
                          labels = c(1, 2, 3))
churn_clean$Gender = factor(churn_clean$Gender,
                            levels = c('Female', 'Male'),
                            labels = c(0, 1))
churn_clean$SeniorCitizen = factor(churn_clean$SeniorCitizen,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$Partner = factor(churn_clean$Partner,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$Dependents = factor(churn_clean$Dependents,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$PhoneService = factor(churn_clean$PhoneService,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$MultipleLines = factor(churn_clean$MultipleLines,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$OnlineSecurity = factor(churn_clean$OnlineSecurity,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$OnlineBackup = factor(churn_clean$OnlineBackup,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$DeviceProtection = factor(churn_clean$DeviceProtection,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$TechSupport = factor(churn_clean$TechSupport,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$StreamingTV = factor(churn_clean$StreamingTV,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$StreamingMovies = factor(churn_clean$StreamingMovies,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$PaperlessBilling = factor(churn_clean$PaperlessBilling,
                            levels = c('No', 'Yes'),
                            labels = c(0, 1))
churn_clean$Churn = factor(churn_clean$Churn,
                            levels = c('No', 'Yes'),
```

```
                              labels = c(0, 1))
#(Eremenko & de Ponteves; Ganesh, 2017)
```

Create Dummy Variables

Dummy variables must be created to represent the categorical variables in the machine learning models.

```
dmy <- dummyVars(" ~ .", data = churn_clean)
churn_clean <- data.frame(predict(dmy, newdata = churn_clean))
#(Amunategui)
```

Train/Test Split

```
library(caTools)
set.seed(123)
split = sample.split(churn_clean$Yes_Churn, SplitRatio = 0.8)
training_set = subset(churn_clean, split == TRUE)
test_set = subset(churn_clean, split == FALSE)
#(Eremenko & de Ponteves)
```

Feature Scaling

Values do not have the same eucledian distance, and therefore do not have the same scale. Transforming values is necessary so all values are on the same scale.

```
training_set[,37] = scale(training_set[,37])
test_set[,37] = scale(test_set[,37])
#(Eremenko & de Ponteves)
```

Export processed data file and save as .xlsx file

```
library(writexl)
library(readxl)

write_xlsx(churn_clean,"Telco_Churn_Clean.xlsx")
```

Mathematical Modeling

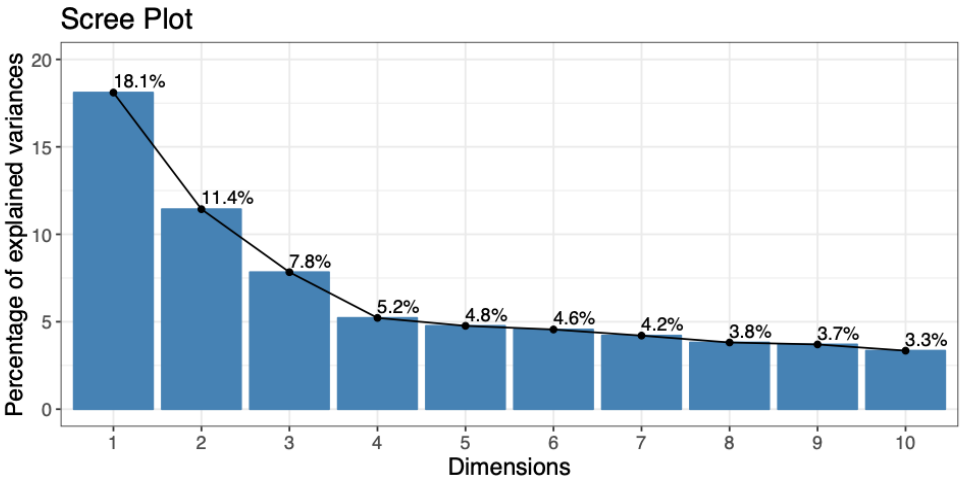
Descriptive Data Mining Method

The main characteristic of descriptive data mining methods is that they lack a dependent variable in the model. These models are often referred to as unsupervised data mining methods. Descriptive data mining methods are used for reducing, summarizing and grouping data. This module describes the most commonly used descriptive data mining methods (Tufféry, 2011).

Principal Component Analysis (PCA)

PCA is one of the most important dimensionality reduction algorithms in machine learning. PCA was used in this analysis to identify the most relevant independent variables in customer churns status (Tufféry, 2011; Kassambara, Fábio, & Visitor, 2017).

```
#PCA Model
library(factoextra)
res.pca <- prcomp(training_set[, 1:42], scale = TRUE)
```

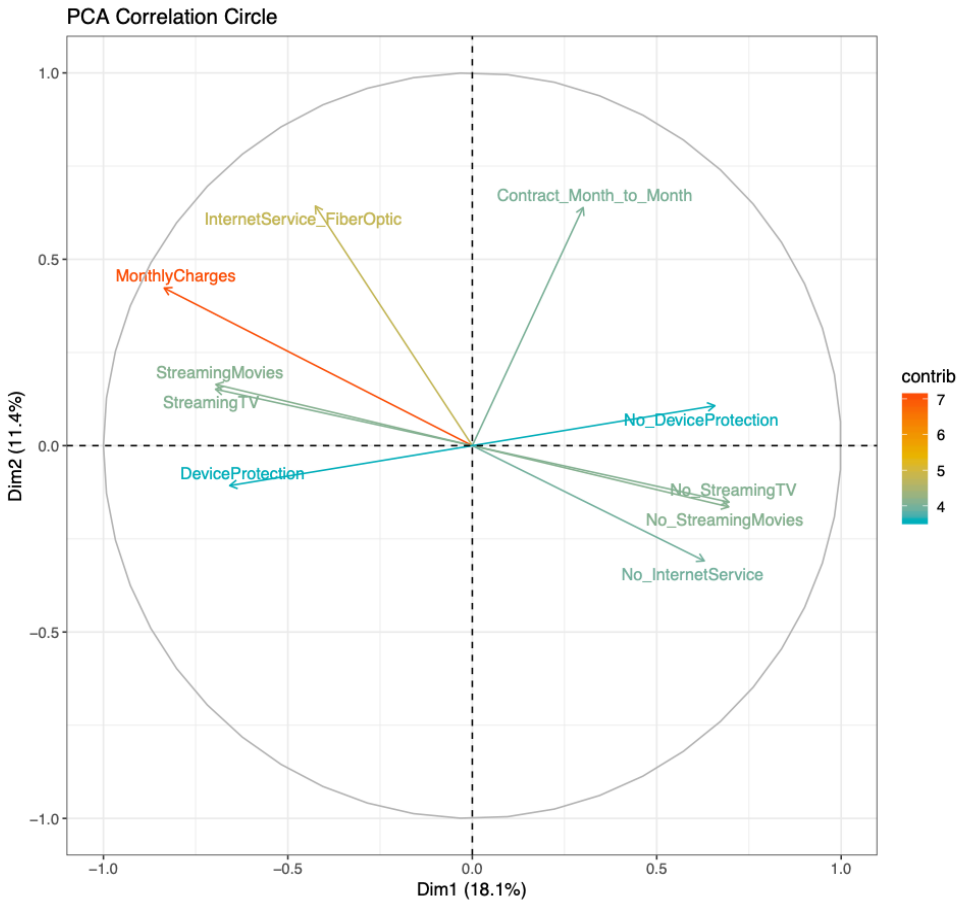


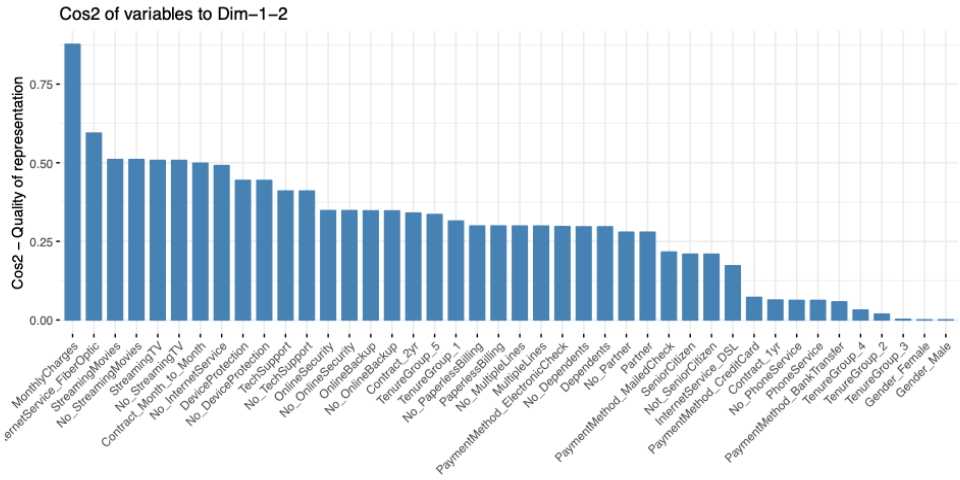
Eigenvalues Table

```
eig.val <- get_eigenvalue(res.pca)
eig.val[1:3,]
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   7.603373        18.103269                      18.10327
## Dim.2   4.802971        11.435645                      29.53891
## Dim.3   3.290349         7.834165                      37.37308
```

PCA results show the first three dimensions account for 37.4% of the explained variance.


PCA Correlation Circle
Contract_Month_to_Month
InternetService_FiberOptic
MonthlyCharges
StreamingMovies
StreamingTV
No_DeviceProtection
DeviceProtection
No_StreamingTV
No_StreamingMovies
No_InternetService
contrib
7
6
5
4
1.0
0.5
0.0
−0.5
−1.0
Dim2 (11.4%)
Dim1 (18.1%)



Note that, a high cos2 value indicates a good representation of the variable on the principal component axis. A variable with a high cos2 value will be represented on the correlation circle by a vector end point near the circumference. A low cos2 indicates that the variable is not perfectly represented by the principal components and will be represented by a short vector close to the circle origin (Kassambara, Fábio, & Visitor, 2017; Tufféry, 2011).

Summary The PCA results show the top variables that influence customer attrition are Monthly Charges, Fiber Optic Internet Service, the ability to stream TV and Movies and to a lesser extent participating in a month-to-month contract.

Predictive Data Mining Methods

Logistic Regression

Logistic Regression is used when the dependent (response) variable is categorical. The primary objective of logistic regression is to model the mean of the response variable, given a set of predictor variables. However, what distinguishes logistic regression from linear regression is that the response variable is binary rather than continuous in nature (Alice, 2015; Eremenko & de Ponteves; Tufféry, 2011).

Train/Test Split Dataframe

To prevent overfitting, the top 8 key performance indicators results from the PCA will be used an input variables for the logistic regression model.

```
#Copy the dataset
churn_ml <- churn_clean
```

```
# Train/Test Split
library(caTools)
set.seed(123)
```

```
split = sample.split(churn_ml$Churn, SplitRatio = 0.8)
training = subset(churn_ml, split == TRUE)
testing = subset(churn_ml, split == FALSE)
```

```
# Feature Scaling
training[,8] = scale(training[,8])
testing[,8] = scale(testing[,8])
```

```
# Fitting Logistic Regression to the Training set
log_classifier <- glm(formula = Churn ~ .,
                      family = binomial,
                      data = training)
summary(log_classifier)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8082  -0.7937  -0.2799   0.6602   3.0134
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.977955   0.126792 -15.600  < 2e-16 ***
## InternetService_FiberOptic      1.098805   0.153915   7.139 9.40e-13 ***
## StreamingTV                     0.315798   0.098289   3.213 0.001314 **
## StreamingMovies                 0.334669   0.097088   3.447 0.000567 ***
## Contract_1yr                   -1.255842   0.110937 -11.320  < 2e-16 ***
## Contract_2yr                   -2.551379   0.183102 -13.934  < 2e-16 ***
## Contract_Month_to_Month              NA         NA      NA       NA
## PaymentMethod_ElectronicCheck  0.503924   0.074829   6.734 1.65e-11 ***
## MonthlyCharges                 -0.009073   0.104340  -0.087 0.930708
## TenureGroup_1                   1.149980   0.084335  13.636  < 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: 6513.9  on 5624  degrees of freedom
## Residual deviance: 4810.4  on 5616  degrees of freedom
## AIC: 4828.4
##
## Number of Fisher Scoring iterations: 6
```

```
#(Eremenko & de Ponteves)
```

Logistic regression results show statistically significant p-values for Fiber Optic Internet, Contracts, and TenureGroup_1. Interestingly, the month-to-contract factor wasn't used in this algorithm due to a high collinearity with another factor. Further classification analysis should be conducted to further identify the importance of this variable (Alice, 2015).

```
#Feature Analysis
anova(log_classifier, test="LRT")
```

```
## 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                                                5624     6513.9
## InternetService_FiberOptic      1   529.16      5623     5984.7 < 2.2e-16
## StreamingTV                     1    11.95      5622     5972.8 0.0005467
## StreamingMovies                 1     3.49      5621     5969.3 0.0617423
## Contract_1yr                    1   173.43      5620     5795.9 < 2.2e-16
## Contract_2yr                    1   716.98      5619     5078.9 < 2.2e-16
## Contract_Month_to_Month         0     0.00      5619     5078.9
## PaymentMethod_ElectronicCheck  1    64.63      5618     5014.3 9.018e-16
## MonthlyCharges                  1     6.68      5617     5007.6 0.0097516
## TenureGroup_1                   1   197.21      5616     4810.4 < 2.2e-16
##
## NULL
## InternetService_FiberOptic    ***
## StreamingTV                   ***
## StreamingMovies               .
## Contract_1yr                  ***
## Contract_2yr                  ***
## Contract_Month_to_Month
## PaymentMethod_ElectronicCheck ***
## MonthlyCharges                **
## TenureGroup_1                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see from the deviance table, the factors that have the greatest impact on churn outcome, in descending order, are Contract_2yr, InternetService_FiberOptic and TenureGroup_1.

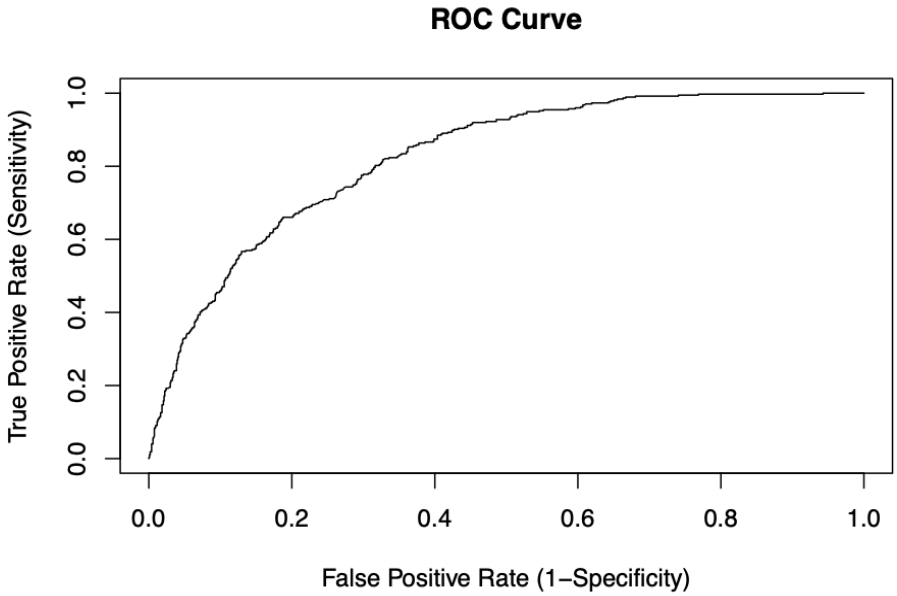
Logistic Regression Predictability Assessment

```
## [1] "Logistic Regression Confusion Matrix"
```

```
##           Actual
## Predicted   0   1
##         0 936 205
##         1  97 169
```

```
## [1] "Logistic Regression Accuracy(%) = 78.5358919687278"
```

A ROC curve and the AUC (area under the curve) will be calculated which are typical performance measurements for a binary classifier. The ROC is a curve created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. A model with good predictive ability should have an AUC closer to 1 (ideal) than to 0.5 (Tufféry, 2011; Vogler, 2016).




```
## [1] "AUC Results = 0.824038028482541"
```

Summary

The Logistic regression algorithm performed well with an accuracy of 78.5% and an AUC value of 0.824. The PCA analysis indicated the streaming media and Monthly Charges were top factors while the logistic regression analysis indicates Tenure and Fiber Optic internet to be of greater importance. Note, Contract_Month_to_Month data was not used in this analysis and the results may be skewed. Future work can be done to enhance the model's accuracy by utilizing a backward elimination method.

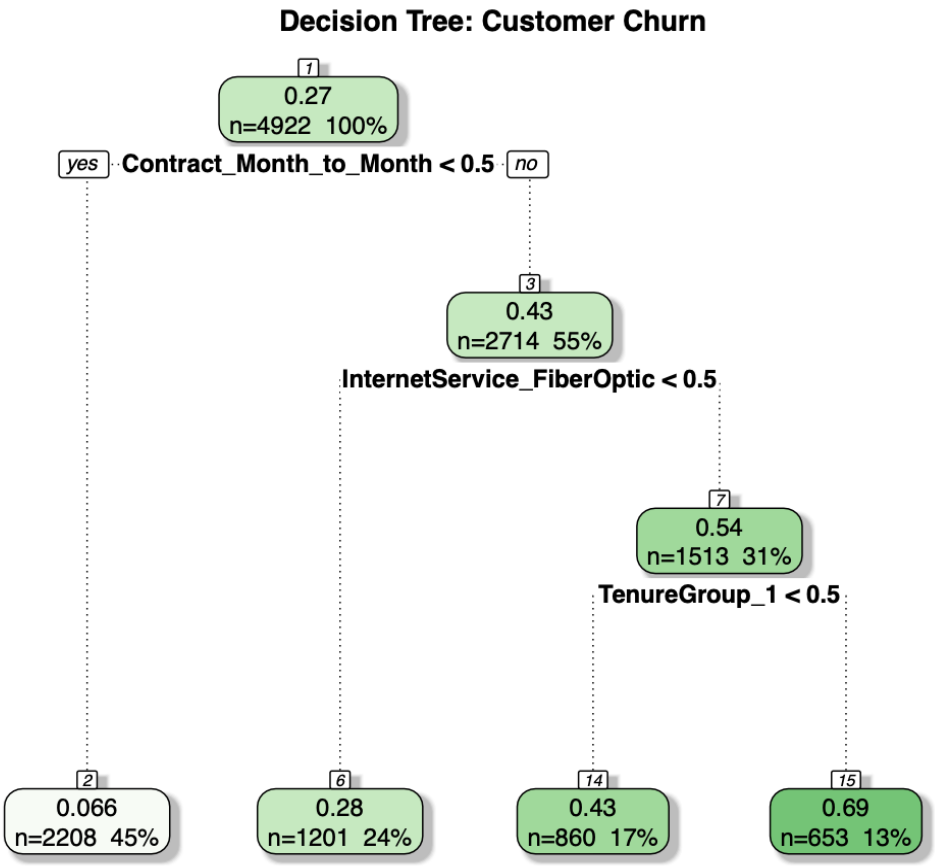
Tree Based Machine Learning Algorithms

Tree based machine learning algorithms are based on the statistical 'bootstrapping' technique. In bootstrapping, given a sample of size N, we create datasets of size N by sampling this original dataset with replacement. Machine Learning models are built on the different bootstrapped samples and then averaged (Rickert, 2013; Tufféry, 2011).

Decision Tree Model

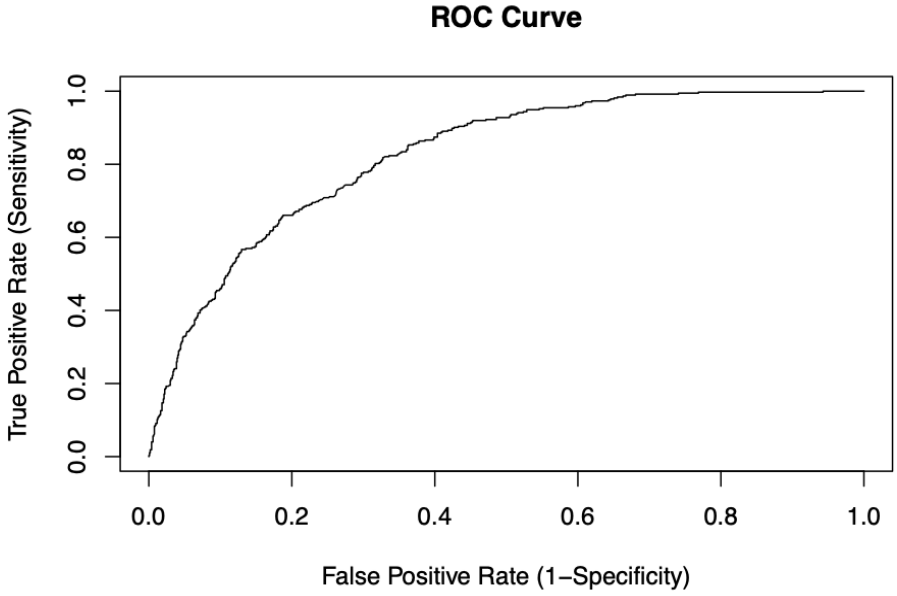
Decision Trees Classifications are a form of regression model in the form of a tree structure. The topmost decision node, known as the "root node", corresponds to the best predictor variable. Decision trees can handle both categorical and numerical data which is the primary reason it is utilized in this data analysis. Using the top-most fluencing variables from the PCA and Logistic Regression results, a Decision Tree analysis will be performed to further identify useful information (Analytics Vidhya Content Team, 2018; Rickert, 2013; Tufféry, 2011).

```
# Fitting Decision Tree to the Training set
tree_classifier <- ctree(Churn~., training)
#(Eremenko & de Ponteves)
```



Decision Tree Predictablity Assessment

```
## [1] "Decision Tree Classification Accuracy(%) = 79.2890995260664"
```




```
## [1] "AUC Results = 0.828119228206571"
```

Summary

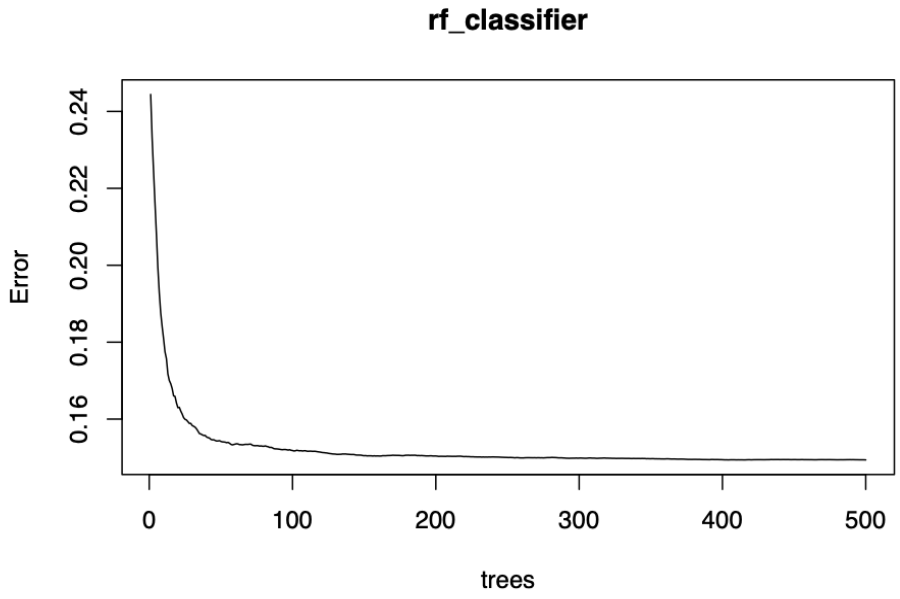
The Decision Tree model has an accuracy of 79.3% and an AUC value equal to 0.83 with results indicating the most important variables in predicting a customer's churn status are month-to-month contract and Fiber Optic Internet Service. A person who is in a month-to-month contract, has fiber optic Internet Service and has been with the company for less than one year, has the highest probability of leaving the company.

Random Forest Classification Model

Random forest is a supervised learning algorithm that builds multiple decision trees and merges them together to get a more accurate and stable prediction (Eremenko & de Ponteves; Ganesh, 2017; Padda, 2018; Rickert, 2013; Tufféry, 2011).

```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training, ntree = 500,      importance = TRUE, proximity =
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 14
##
##           Mean of squared residuals: 0.1493893
##                     % Var explained: 23.44
```

Out of Bag (OOB) Error: In Random Forest and Gradient Boosting for each bootstrap sample taken from the dataset, there will be samples left out, these are known as Out of Bag samples. Classification accuracy carried out on these OOB samples is known as OOB error (Padda, 2018; Rickert, 2013; Tufféry, 2011).

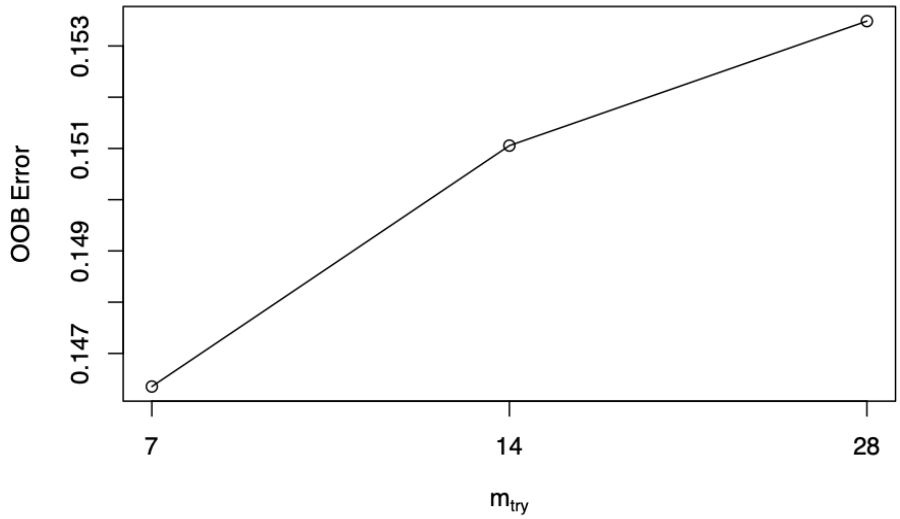


Random Forest Predictablity Assessment

```
## [1] "Random Forest Classification Accuracy(%) = 78.7203791469194"
```

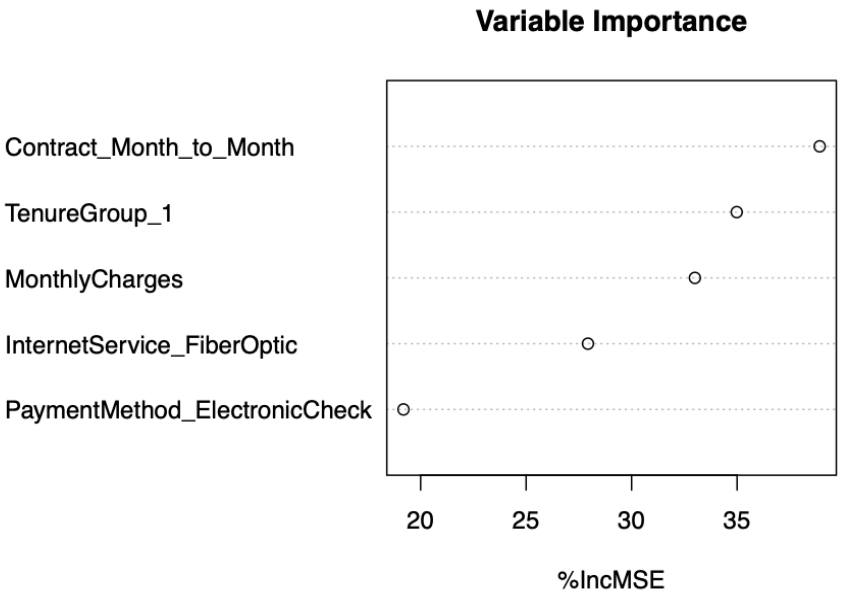
```
#Tuning
rf_tune <- tuneRF(training[, -43],
                  training[, 43],
                  stepFactor = 0.5,
                  plot = TRUE,
                  ntreeTry = 150,
                  trace = TRUE,
                  improve = 0.05)
```

```
## mtry = 14  OOB error = 0.1510553
## Searching left ...
## mtry = 28    OOB error = 0.153485
## -0.01608512 0.05
## Searching right ...
## mtry = 7     OOB error = 0.1463533
## 0.0311276 0.05
```



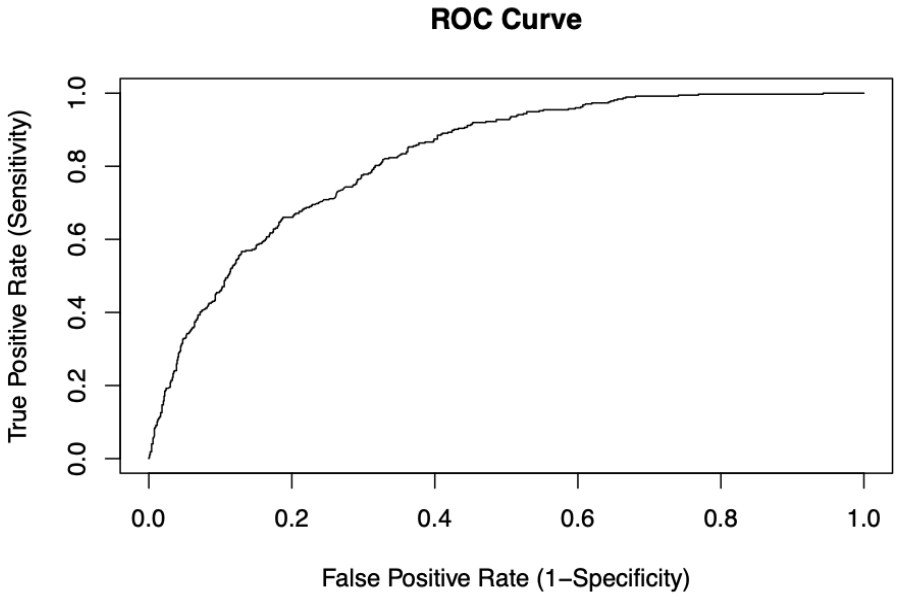

```
#(Eremenko & de Ponteves; Ganesh, 2017; Padda, 2018; Rickert, 2013)
```

```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training, ntree = 100,      mtry = 7, importance = TRUE, p:
##                Type of random forest: regression
##                      Number of trees: 100
## No. of variables tried at each split: 7
##
##           Mean of squared residuals: 0.1471876
##                     % Var explained: 24.57
```



The percent increase in mean square error (MSE) for each variable if it were omitted from the analysis. Therefore, the higher the %IncMSE value, the more important the variable in determing outcome (Padda, 2018; Rickert, 2013; Tufféry, 2011).

```
## [1] "Random Forest Classification Accuracy(%) = 78.6255924170616"
```




```
## [1] "AUC Results = 0.817550624921607"
```

The Random Forest machine learning algorithm performed at a 78.8% accuracy.Interestingly, the Random Forest results show the top most important independent variables to be a month-to-month contract, Tenure less than 1 year and the total monthly charges which differs from the Decision Tree model results. Future work to improve upon the accuracy of the Random Forest can be done by creating a random forest iteration with fewer trees and include fewer important variables.

Conclusion

An analysis on customer churn data was conducted to better predict customer churn outcomes. Customer attrition, also known as churn, is defined to be when customers stop doing business with a company. Churn is a growth decelerator and tracking the churn rate is essential to business success (Customer Churn Prediction, Prevention & Analysis, 2017; Tufféry, 2011). The telecommunications dataset contained 7043 observations with 20 independent variables and a bininary, dependant variable with the customer churn outcome. The dataset was imported into R Studio using the readr library. Data preprocessing techniques were used to address qualify and tidyness issues. A Univariate analysis was conducted on the 17 discrete categorical variables. Bar charts were utilized to display the frequency data. Discriptive analysis show that the dataset has an even distribution of men and women. The majority of customers have been with company less than a year, participate in paperless billing, have a month-to-month contract and pay their bill using an electonic check. A Pearson correlation coefficient was computed to assess the statistical relationship between the Monthly and Total charges billed to customers. Increases in monthly charges were moderately correlated,r = 0.65, with increases in total charges. Due to the correlative relationship between the variables, the Total Charges factor was removed from the dataset to help model accuracy.

Several statistical algorithms were used to understand the relationship between the independent variables and the target variable. Principal component analysis (PCA), a dimension-reducing technique, was used to

identify the top-most relevant independent variables in customer churns status. Logistic Regression models are ideal for binary, categorical analysis and was used in this analysis. To prevent overfitting, the top 8 key performance indicator results from the PCA were used in the logistic regression model. Tree based learning algorithms were used in this analysis because they are considered to be one of the best and mostly used supervised learning methods because they empower predictive models with high accuracy, stability and ease of interpretation (Analytics Vidhya Content Team, 2018).

The PCA analysis showed the top performance indictors in determining customer churn are monthly charges, fiber optic internet and month-to-month contracts. The logistic regression analysis results show Tenure Less than 1 year and Fiber Optic internet to be of the key performance indicators for predicting attrition. The Decision Tree model had the best accuracy result, showing Month-to_Month Contract, Fiber Optic Internet Service, and Tenure (less than 1 year) as the key factors customer churn. Interestingly, the Random Forest model performed comparatively to the Logistic Regression Model in terms of accuracy and AUC performance. The results of the Random Forest model also found Month-to-Month Contract as the top most factor, and placed Tenure (less than 1 year) and Monthly Charges over Fiber Optic Internet Service in order of variable importance.

In summary, key indicator variables associated with customer churn are Month-to_Month Contract, Monthly Charges,Fiber Optic Internet Service and Tenure (less than 1 year). Results show a high churn probability in customers who are in a month-to-month contract, have been with the company less than one year and have Fiber Optic internet service. Customers with the lowest churn probability participate in 1- or 2-year contracts, have DSL internet and have been with the company more than one year and do not participate in paperless billing.

References

1. Analytics Vidhya Content Team. (2018, April 18). A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python). Retrieved from https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree- based-modeling-scratch-in-python/
2. Alice, M. (2015, September 13). How to perform a Logistic Regression in R. Retrieved from https://www.r-bloggers.com/how-to-perform-a-logistic -regression-in-r/ (Alice, 2015).
3. Amunategui, M. (n.d.). Brief Walkthrough Of The dummyVars Function From {caret}. Retrieved December 12, 2018, from https://amunategui.github. io/dummyVar-Walkthrough/
4. Condition a ..count.. summation on the faceting variable. (2012, July 19). Retrieved from https://stackoverflow.com/questions/11567389/condition- a-count-summation-on-the-faceting-variable
5. Correlation matrix: A quick start guide to analyze, format and visualize a correlation matrix using R software. (n.d.). Retrieved December 12, 2018 from http://www.sthda.com/english/wiki/correlation-matrix-a-quick -start-guide-to-analyze-format-and-visualize-a-correlation-matrix- using-r-software
6. Customer Churn Prediction, Prevention & Analysis. (2017, August). Retrieved December 1, 2018, from https://www.optimove.com/learning- center/customer-churn-prediction-and-prevention
7. Eremenko, K., & De Ponteves, H. (n.d.). Machine Learning A-Z (Python & R in Data Science Course). Retrieved December 1, 2018, from https://www.udemy.com/machinelearning
8. Ganesh, T. V. (2017, November 06). Practical Machine Learning with R and Python – Part 5. Retrieved from https://www.r-bloggers.com/practical- machine-learning-with-r-and-python-part-5/
9. Kassambara, Fábio, & Visitor. (2017, September 24). CA - Correspondence Analysis in R: Essentials. Retrieved from http://www.sthda.com/english/articles/31-principal-component-methods- in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/
10. Padda, A. (2018, January 26). Introduction to Random Forest. Retrieved from https://analyticsdefined.com/introduction-random-forests/

11. Prone-R, D. (2016, February 16). Multiple regression lines in ggpairs. Retrieved from https://www.r-bloggers.com/multiple-regression-lines- in-ggpairs/

12. R Documentation: Levels Attributes. (n.d.). Retrieved December 12, 2018, from https://stat.ethz.ch/R-manual/R-devel/library/base/html/levels.html

13. Rickert, J. (2013, June 19). Draw nicer Classification and Regression Trees with the rpart.plot package. Retrieved from https://blog.revolutionanalytics.com/2013/06/plotting-classification- and-regression-trees-with-plotrpart.html

14. Rokicki, S. (2012, October 22). From continuous to categorical. Retrieved from https://www.r-bloggers.com/from-continuous-to-categorical/

15. Statistical tools for high-throughput data analysis (sthda). (n.d.). Add titles to a plot in R software. Retrieved from http://www.sthda.com/ english/wiki/add-titles-to-a-plot-in-r-software

16. Tuffery, S. (2011). Data mining and statistics for decision making (1st Edition ed.). [Western Governors University]. Retrieved from https://wgu.vitalsource.com/#/books/undefined

17. Vogler, R. (2016, June 11). Illustrated Guide to ROC and AUC. Retrieved from https://www.r-bloggers.com/illustrated-guide-to-roc-and-auc/

18. Wingate, R. (2018, July 17). Data Quality & Tidiness. Retrieved from December 1, 2018, from https://ryanwingate.com/purpose/tidy-data/