Model Results

### Customer Churn Prediction

Data mining models Created : 12/1/2017 Modified : 12/14/2017

References : <https://datascienceplus.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/>

Customer churn is when customers or subscribers to services of a company quit using their services or stip doing business with the company. This analysis uses the telecom churn dataset from IBM Watson Analytics. The business challenge for this problem is - A telecommunications company [Telco] is concerned about the number of customers leaving their landline business for cable competitors. They need to understand who is leaving. Imagine that you’re an analyst at this company and you have to find out who is leaving and why.

rm(list = ls(all = TRUE))  
  
## Load the necessary packages  
library(dplyr)  
library(corrplot)  
library(ggplot2)  
library(gridExtra)  
library(ggthemes)  
library(MASS)  
library(caret)  
library(randomForest)  
library(party)  
library(summarytools)

###=====================================================#  
### Data Extraction and Cleaning  
###=====================================================#  
  
## Read in the data file and save it   
# fileLoc <- "https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA\_Fn-UseC\_-Telco-Customer-Churn.csv?cm\_mc\_uid=60538928144815121676500&cm\_mc\_sid\_50200000=1512167650&cm\_mc\_sid\_52640000=1512167650"  
  
# churn <- read.csv(fileLoc)  
# save(churn, file = "churn.RData")  
  
load("churn.RData")  
str(churn)

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...

Data includes customer Id, demographics, payment etc. There are 7043 observations and 21 variables. The arget variable is Churn. Churn - customers who left within the last month Services - phone, multiple lines, internet, online security, online backup, device protection, tech support and streaming movies Account Info. - tenure, contract, payment method, paperless billing, monthly charges and total charges. Demographics - Gender, age range, and if they had partners or dependents

colSums(is.na(churn))

## 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

## There are 11 obs with NAs in TotalCharges. Removing these rows.  
  
churn <- churn[complete.cases(churn), ]  
  
head(churn)

## customerID gender SeniorCitizen Partner Dependents tenure PhoneService  
## 1 7590-VHVEG Female 0 Yes No 1 No  
## 2 5575-GNVDE Male 0 No No 34 Yes  
## 3 3668-QPYBK Male 0 No No 2 Yes  
## 4 7795-CFOCW Male 0 No No 45 No  
## 5 9237-HQITU Female 0 No No 2 Yes  
## 6 9305-CDSKC Female 0 No No 8 Yes  
## MultipleLines InternetService OnlineSecurity OnlineBackup  
## 1 No phone service DSL No Yes  
## 2 No DSL Yes No  
## 3 No DSL Yes Yes  
## 4 No phone service DSL Yes No  
## 5 No Fiber optic No No  
## 6 Yes Fiber optic No No  
## DeviceProtection TechSupport StreamingTV StreamingMovies Contract  
## 1 No No No No Month-to-month  
## 2 Yes No No No One year  
## 3 No No No No Month-to-month  
## 4 Yes Yes No No One year  
## 5 No No No No Month-to-month  
## 6 Yes No Yes Yes Month-to-month  
## PaperlessBilling PaymentMethod MonthlyCharges TotalCharges  
## 1 Yes Electronic check 29.85 29.85  
## 2 No Mailed check 56.95 1889.50  
## 3 Yes Mailed check 53.85 108.15  
## 4 No Bank transfer (automatic) 42.30 1840.75  
## 5 Yes Electronic check 70.70 151.65  
## 6 Yes Electronic check 99.65 820.50  
## Churn  
## 1 No  
## 2 No  
## 3 Yes  
## 4 No  
## 5 Yes  
## 6 Yes

dfSummary(churn, style = "grid", plain.ascii = TRUE)

##   
## Dataframe Summary  
##   
## churn  
##   
##   
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | Variable | Properties | Stats / Values | Freqs, % Valid | N Valid |  
## +==================+===============+==========================================+======================+===========+  
## | customerID | type:integer | 1. 0002-ORFBO | 1: 1 (0%) | 7032/7032 |  
## | | class:factor | 2. 0003-MKNFE | 2: 1 (0%) | (100.0%) |  
## | | | 3. 0004-TLHLJ | 3: 1 (0%) | |  
## | | | 4. 0011-IGKFF | 4: 1 (0%) | |  
## | | | 5. 0013-EXCHZ | 5: 1 (0%) | |  
## | | | 6. 0013-MHZWF | 6: 1 (0%) | |  
## | | | 7. 0013-SMEOE | 7: 1 (0%) | |  
## | | | 8. 0014-BMAQU | 8: 1 (0%) | |  
## | | | 9. 0015-UOCOJ | 9: 1 (0%) | |  
## | | | 10. 0016-QLJIS | 10: 1 (0%) | |  
## | | | ... 7033 other levels | others: 7022 (99.9%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | gender | type:integer | 1. Female | 1: 3483 (49.5%) | 7032/7032 |  
## | | class:factor | 2. Male | 2: 3549 (50.5%) | (100.0%) |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | SeniorCitizen | type:integer | mean (sd) = 0.16 (0.37) | 0: 5890 (83.8%) | 7032/7032 |  
## | | class:integer | min < med < max = 0 < 0 < 1 | 1: 1142 (16.2%) | (100.0%) |  
## | | | IQR (CV) = 0 (2.27) | | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | Partner | type:integer | 1. No | 1: 3639 (51.7%) | 7032/7032 |  
## | | class:factor | 2. Yes | 2: 3393 (48.3%) | (100.0%) |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | Dependents | type:integer | 1. No | 1: 4933 (70.2%) | 7032/7032 |  
## | | class:factor | 2. Yes | 2: 2099 (29.8%) | (100.0%) |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | tenure | type:integer | mean (sd) = 32.42 (24.55) | 72 distinct values | 7032/7032 |  
## | | class:integer | min < med < max = 1 < 29 < 72 | | (100.0%) |  
## | | | IQR (CV) = 46 (0.76) | | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | PhoneService | type:integer | 1. No | 1: 680 (9.7%) | 7032/7032 |  
## | | class:factor | 2. Yes | 2: 6352 (90.3%) | (100.0%) |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | MultipleLines | type:integer | 1. No | 1: 3385 (48.1%) | 7032/7032 |  
## | | class:factor | 2. No phone service | 2: 680 (9.7%) | (100.0%) |  
## | | | 3. Yes | 3: 2967 (42.2%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | InternetService | type:integer | 1. DSL | 1: 2416 (34.4%) | 7032/7032 |  
## | | class:factor | 2. Fiber optic | 2: 3096 (44%) | (100.0%) |  
## | | | 3. No | 3: 1520 (21.6%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | OnlineSecurity | type:integer | 1. No | 1: 3497 (49.7%) | 7032/7032 |  
## | | class:factor | 2. No internet service | 2: 1520 (21.6%) | (100.0%) |  
## | | | 3. Yes | 3: 2015 (28.7%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | OnlineBackup | type:integer | 1. No | 1: 3087 (43.9%) | 7032/7032 |  
## | | class:factor | 2. No internet service | 2: 1520 (21.6%) | (100.0%) |  
## | | | 3. Yes | 3: 2425 (34.5%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | DeviceProtection | type:integer | 1. No | 1: 3094 (44%) | 7032/7032 |  
## | | class:factor | 2. No internet service | 2: 1520 (21.6%) | (100.0%) |  
## | | | 3. Yes | 3: 2418 (34.4%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | TechSupport | type:integer | 1. No | 1: 3472 (49.4%) | 7032/7032 |  
## | | class:factor | 2. No internet service | 2: 1520 (21.6%) | (100.0%) |  
## | | | 3. Yes | 3: 2040 (29%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | StreamingTV | type:integer | 1. No | 1: 2809 (39.9%) | 7032/7032 |  
## | | class:factor | 2. No internet service | 2: 1520 (21.6%) | (100.0%) |  
## | | | 3. Yes | 3: 2703 (38.4%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | StreamingMovies | type:integer | 1. No | 1: 2781 (39.5%) | 7032/7032 |  
## | | class:factor | 2. No internet service | 2: 1520 (21.6%) | (100.0%) |  
## | | | 3. Yes | 3: 2731 (38.8%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | Contract | type:integer | 1. Month-to-month | 1: 3875 (55.1%) | 7032/7032 |  
## | | class:factor | 2. One year | 2: 1472 (20.9%) | (100.0%) |  
## | | | 3. Two year | 3: 1685 (24%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | PaperlessBilling | type:integer | 1. No | 1: 2864 (40.7%) | 7032/7032 |  
## | | class:factor | 2. Yes | 2: 4168 (59.3%) | (100.0%) |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | PaymentMethod | type:integer | 1. Bank transfer (automatic) | 1: 1542 (21.9%) | 7032/7032 |  
## | | class:factor | 2. Credit card (automatic) | 2: 1521 (21.6%) | (100.0%) |  
## | | | 3. Electronic check | 3: 2365 (33.6%) | |  
## | | | 4. Mailed check | 4: 1604 (22.8%) | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | MonthlyCharges | type:double | mean (sd) = 64.8 (30.09) | 1584 distinct values | 7032/7032 |  
## | | class:numeric | min < med < max = 18.25 < 70.35 < 118.75 | | (100.0%) |  
## | | | IQR (CV) = 54.27 (0.46) | | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | TotalCharges | type:double | mean (sd) = 2283.3 (2266.77) | 6530 distinct values | 7032/7032 |  
## | | class:numeric | min < med < max = 18.8 < 1397.47 < | | (100.0%) |  
## | | | 8684.8 | | |  
## | | | IQR (CV) = 3393.29 (0.99) | | |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+  
## | Churn | type:integer | 1. No | 1: 5163 (73.4%) | 7032/7032 |  
## | | class:factor | 2. Yes | 2: 1869 (26.6%) | (100.0%) |  
## +------------------+---------------+------------------------------------------+----------------------+-----------+

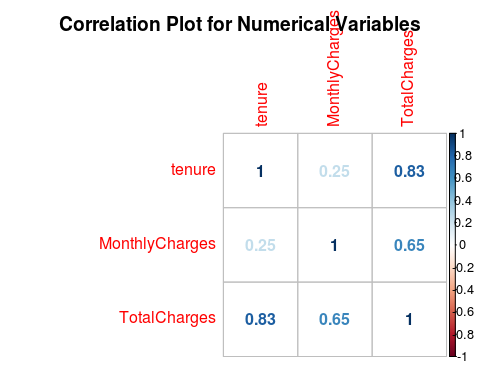
## As evident in the summary, variables needs cleaning up. For e.g. convert "No internet service" to "No".  
churn <- churn %>%   
 mutate(MultipleLines = ifelse(MultipleLines %in% c("No", "No phone service"), "No", "Yes"),  
 OnlineSecurity = ifelse(OnlineSecurity %in% c("No", "No internet service"), "No", "Yes"),  
 OnlineBackup = ifelse(OnlineBackup %in% c("No", "No internet service"), "No", "Yes"),  
 DeviceProtection = ifelse(DeviceProtection %in% c("No", "No internet service"), "No", "Yes"),  
 TechSupport = ifelse(TechSupport %in% c("No", "No internet service"), "No", "Yes"),  
 StreamingTV = ifelse(StreamingTV %in% c("No", "No internet service"), "No", "Yes"),  
 StreamingMovies = ifelse(StreamingMovies %in% c("No", "No internet service"), "No", "Yes"),  
 SeniorCitizen = ifelse(SeniorCitizen == 1, "Yes", "No"))  
  
  
  
## Range of tenure is between 1 and 72  
range(churn$tenure)

## [1] 1 72

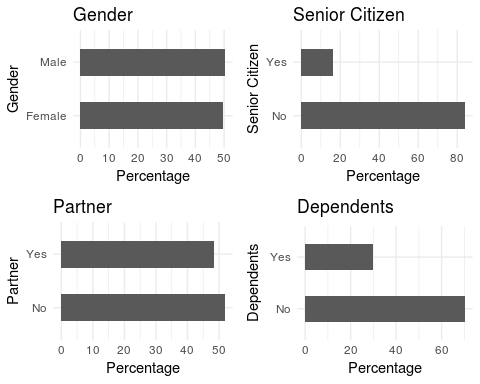
churn <- churn %>%   
 mutate(tenure\_group = cut(tenure,   
 breaks = c(0, 12, 24, 48, 60, Inf),   
 labels = c("0-12 Month", "12-24 Month", "24-48 Month", "48-60 Month", ">60 Month")))  
table(churn$tenure\_group)

##   
## 0-12 Month 12-24 Month 24-48 Month 48-60 Month >60 Month   
## 2175 1024 1594 832 1407

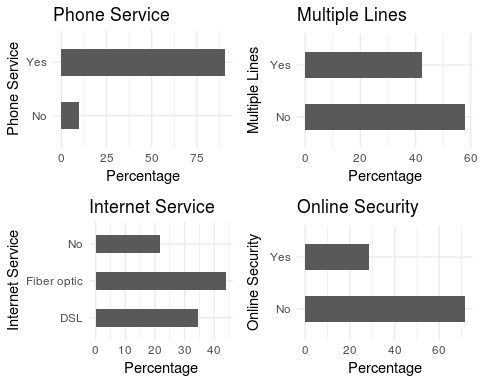
## Convert modified variables from character to factors  
columns <- c("MultipleLines", "OnlineSecurity", "OnlineBackup", "DeviceProtection",  
 "TechSupport", "StreamingTV", "StreamingMovies", "SeniorCitizen", "tenure\_group")  
   
churn[columns] <- lapply(churn[columns], factor)  
  
###========================================================#  
## Exploratory Data Analysis  
###========================================================#  
  
## Correlation between numeric terms  
numeric.var <- sapply(churn, is.numeric)  
corr.matrix <- cor(churn[,numeric.var])  
corrplot(corr.matrix, main="\n\nCorrelation Plot for Numerical Variables", method="number")



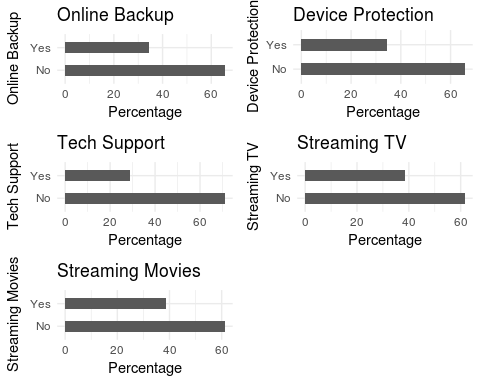
## Since monthly charges and total charges are correlated, one of them is removed.  
  
## Finally, remove variable not required for modeling   
churn <- churn %>%   
 dplyr::select(-c(TotalCharges, tenure, customerID))  
  
##==========================  
## Visualize the attributes  
##==========================  
  
drawPlot <- function(var, xlab){  
 sub.churn <- as.data.frame(churn[,var])  
 names(sub.churn) <- "category"  
 p <- ggplot(sub.churn, aes(x=category)) + ggtitle(xlab) + xlab(xlab) +  
 geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()  
}  
  
  
## Demographics  
p1 <- drawPlot("gender", "Gender")  
p2 <- drawPlot("SeniorCitizen", "Senior Citizen")  
p3 <- drawPlot("Partner", "Partner")  
p4 <- drawPlot("Dependents", "Dependents")  
grid.arrange(p1, p2, p3, p4, ncol=2)



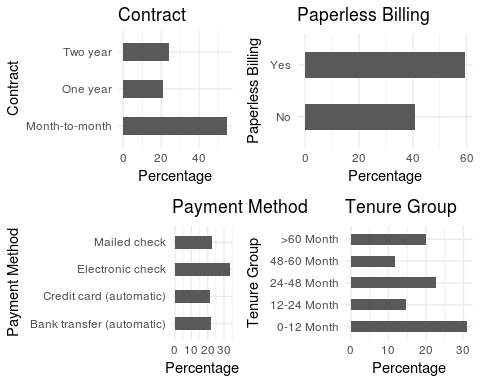
## Balanced male-female ratio of customers. Majority of whom are non senior citizens and independent.   
  
## Visualize service availability  
p5 <- drawPlot("PhoneService", "Phone Service")  
p6 <- drawPlot("MultipleLines", "Multiple Lines")  
p7 <- drawPlot("InternetService", "Internet Service")  
p8 <- drawPlot("OnlineSecurity", "Online Security")  
grid.arrange(p5, p6, p7, p8, ncol=2)



## Majority of customers have phone service, multiple lines, online security, fiber optic internet service.  
  
## Visualize other services  
p9 <- drawPlot("OnlineBackup", "Online Backup")  
p10 <- drawPlot("DeviceProtection", "Device Protection")  
p11 <- drawPlot("TechSupport", "Tech Support")  
p12 <- drawPlot("StreamingTV", "Streaming TV")  
p13 <- drawPlot("StreamingMovies", "Streaming Movies")  
grid.arrange(p9, p10, p11, p12, p13, ncol=2)



## Also the majority of customers have the following features - online backup, device protection, tech support, streaming TV and streaming movies  
  
## Payment features  
p14 <- drawPlot("Contract", "Contract")  
p15 <- drawPlot("PaperlessBilling", "Paperless Billing")  
p16 <- drawPlot("PaymentMethod", "Payment Method")  
p17 <- drawPlot("tenure\_group", "Tenure Group")   
grid.arrange(p14, p15, p16, p17, ncol=2)



## Majority of customers have month-to-month contract, paperless billing, and pay electronic checks. Finally, majority of the customers are in within the 1 year tenure.  
  
## The data now has all the bunch of variables relating to the features of an individual customer.  
  
###===================================================#  
## Mining Models  
###===================================================#  
  
##=============================  
## 1. Logistic Regression  
##=============================  
  
## Create test and training set  
set.seed(2017)  
intrain <- createDataPartition(churn$Churn, p = 0.7, list = FALSE)  
  
train <- churn[intrain, ]  
test <- churn[-intrain, ]  
  
dim(train); dim(test)

## [1] 4924 19

## [1] 2108 19

## Fit the Logistic Regression Model  
glm.fit <- glm(Churn ~ ., family = binomial(link = "logit"), data = train)  
print(summary(glm.fit))

##   
## Call:  
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0642 -0.6579 -0.2841 0.6647 3.0802   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.542936 0.989692 0.549 0.58329  
## genderMale 0.005215 0.078054 0.067 0.94673  
## SeniorCitizenYes 0.268218 0.100590 2.666 0.00767  
## PartnerYes -0.031401 0.094273 -0.333 0.73907  
## DependentsYes -0.126373 0.108965 -1.160 0.24615  
## PhoneServiceYes 0.039976 0.796526 0.050 0.95997  
## MultipleLinesYes 0.395405 0.217701 1.816 0.06933  
## InternetServiceFiber optic 1.568899 0.977972 1.604 0.10866  
## InternetServiceNo -1.445689 0.992016 -1.457 0.14503  
## OnlineSecurityYes -0.215432 0.218837 -0.984 0.32490  
## OnlineBackupYes -0.061062 0.213932 -0.285 0.77532  
## DeviceProtectionYes 0.018193 0.215515 0.084 0.93272  
## TechSupportYes -0.210097 0.220270 -0.954 0.34018  
## StreamingTVYes 0.575121 0.398781 1.442 0.14925  
## StreamingMoviesYes 0.580613 0.400970 1.448 0.14761  
## ContractOne year -0.884153 0.132181 -6.689 2.25e-11  
## ContractTwo year -1.625461 0.212907 -7.635 2.27e-14  
## PaperlessBillingYes 0.274043 0.089526 3.061 0.00221  
## PaymentMethodCredit card (automatic) -0.222822 0.136348 -1.634 0.10221  
## PaymentMethodElectronic check 0.175685 0.112318 1.564 0.11778  
## PaymentMethodMailed check -0.108172 0.136507 -0.792 0.42811  
## MonthlyCharges -0.025536 0.038938 -0.656 0.51195  
## tenure\_group12-24 Month -0.900290 0.117629 -7.654 1.95e-14  
## tenure\_group24-48 Month -1.353930 0.120910 -11.198 < 2e-16  
## tenure\_group48-60 Month -1.583778 0.169371 -9.351 < 2e-16  
## tenure\_group>60 Month -1.788663 0.204671 -8.739 < 2e-16  
##   
## (Intercept)   
## genderMale   
## SeniorCitizenYes \*\*   
## PartnerYes   
## DependentsYes   
## PhoneServiceYes   
## MultipleLinesYes .   
## InternetServiceFiber optic   
## InternetServiceNo   
## OnlineSecurityYes   
## OnlineBackupYes   
## DeviceProtectionYes   
## TechSupportYes   
## StreamingTVYes   
## StreamingMoviesYes   
## ContractOne year \*\*\*  
## ContractTwo year \*\*\*  
## PaperlessBillingYes \*\*   
## PaymentMethodCredit card (automatic)   
## PaymentMethodElectronic check   
## PaymentMethodMailed check   
## MonthlyCharges   
## tenure\_group12-24 Month \*\*\*  
## tenure\_group24-48 Month \*\*\*  
## tenure\_group48-60 Month \*\*\*  
## tenure\_group>60 Month \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5702.8 on 4923 degrees of freedom  
## Residual deviance: 4053.3 on 4898 degrees of freedom  
## AIC: 4105.3  
##   
## Number of Fisher Scoring iterations: 6

## Odds ratio  
exp(cbind(OR=coef(glm.fit), confint(glm.fit)))

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 1.7210525 0.24775679 12.0054614  
## genderMale 1.0052289 0.86262195 1.1714566  
## SeniorCitizenYes 1.3076321 1.07349443 1.5925410  
## PartnerYes 0.9690866 0.80564962 1.1659492  
## DependentsYes 0.8812861 0.71128542 1.0904732  
## PhoneServiceYes 1.0407857 0.21854331 4.9652449  
## MultipleLinesYes 1.4849855 0.96955965 2.2766575  
## InternetServiceFiber optic 4.8013595 0.70783924 32.7586312  
## InternetServiceNo 0.2355838 0.03363247 1.6446115  
## OnlineSecurityYes 0.8061930 0.52477012 1.2377357  
## OnlineBackupYes 0.9407647 0.61847682 1.4309598  
## DeviceProtectionYes 1.0183598 0.66746672 1.5539331  
## TechSupportYes 0.8105058 0.52603764 1.2477392  
## StreamingTVYes 1.7773458 0.81407027 3.8882009  
## StreamingMoviesYes 1.7871332 0.81513626 3.9268353  
## ContractOne year 0.4130639 0.31768593 0.5335691  
## ContractTwo year 0.1968209 0.12803628 0.2954498  
## PaperlessBillingYes 1.3152712 1.10386985 1.5680850  
## PaymentMethodCredit card (automatic) 0.8002569 0.61216278 1.0449771  
## PaymentMethodElectronic check 1.1920626 0.95697593 1.4865810  
## PaymentMethodMailed check 0.8974736 0.68690462 1.1731963  
## MonthlyCharges 0.9747876 0.90306955 1.0520314  
## tenure\_group12-24 Month 0.4064517 0.32229461 0.5111763  
## tenure\_group24-48 Month 0.2582235 0.20337177 0.3267314  
## tenure\_group48-60 Month 0.2051984 0.14667829 0.2850290  
## tenure\_group>60 Month 0.1671836 0.11144473 0.2487311

## Interpretation of results  
## Males have higher odds to churn than females  
  
anova(glm.fit, test = "Chisq")

## 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 4923 5702.8   
## gender 1 0.29 4922 5702.5 0.591863   
## SeniorCitizen 1 109.58 4921 5592.9 < 2.2e-16 \*\*\*  
## Partner 1 128.30 4920 5464.6 < 2.2e-16 \*\*\*  
## Dependents 1 29.67 4919 5434.9 5.118e-08 \*\*\*  
## PhoneService 1 1.31 4918 5433.6 0.252861   
## MultipleLines 1 2.67 4917 5430.9 0.102501   
## InternetService 2 477.50 4915 4953.4 < 2.2e-16 \*\*\*  
## OnlineSecurity 1 168.33 4914 4785.1 < 2.2e-16 \*\*\*  
## OnlineBackup 1 79.43 4913 4705.7 < 2.2e-16 \*\*\*  
## DeviceProtection 1 57.82 4912 4647.9 2.874e-14 \*\*\*  
## TechSupport 1 76.99 4911 4570.9 < 2.2e-16 \*\*\*  
## StreamingTV 1 3.41 4910 4567.5 0.064829 .   
## StreamingMovies 1 1.19 4909 4566.3 0.275058   
## Contract 2 304.03 4907 4262.2 < 2.2e-16 \*\*\*  
## PaperlessBilling 1 8.94 4906 4253.3 0.002785 \*\*   
## PaymentMethod 3 26.64 4903 4226.7 6.999e-06 \*\*\*  
## MonthlyCharges 1 0.78 4902 4225.9 0.375784   
## tenure\_group 4 172.54 4898 4053.3 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of deviance table : https://stats.stackexchange.com/questions/59879/logistic-regression-anova-chi-square-test-vs-significance-of-coefficients-ano  
## https://rstudio-pubs-static.s3.amazonaws.com/108528\_3395f9cf41c04335aa0b1b291e8de72e.html  
  
## The summary table gives the test of hypothesis of comparing a smaller model to an incrmental model starting from first variable to last. The deviance is the reduction after adding the new variable.  
## As observed, adding "InterNetService", "Contract" and "tenure\_group" are the top three variables resulting in a drop in deviance. Although other variables too are statistically significant, the effect size is lower.  
  
## Use this model to predict churn  
test$Churn1 <- as.character(test$Churn)  
test$Churn1[test$Churn1 == "No"] <- "0"  
test$Churn1[test$Churn1 == "Yes"] <- "1"  
  
glm.probs <- predict(glm.fit, newdata = test, type = "response")  
glm.pred <- rep("0", length(glm.probs))  
glm.pred[glm.probs > 0.5] <- "1"  
  
table(glm.pred, test$Churn1)

##   
## glm.pred 0 1  
## 0 1402 298  
## 1 146 262

## Function to compute accuracy, precision and recall  
## https://stackoverflow.com/questions/12572357/precision-recall-and-f-measure-in-r  
measurePrecisionRecall <- function(predict, actual\_labels){  
 precision <- sum(predict & actual\_labels) / sum(predict)  
 recall <- sum(predict & actual\_labels) / sum(actual\_labels)  
 fmeasure <- 2 \* precision \* recall / (precision + recall)  
   
 cat('precision: ')  
 cat(precision \* 100)  
 cat('%')  
 cat('\n')  
   
 cat('recall: ')  
 cat(recall \* 100)  
 cat('%')  
 cat('\n')  
   
 cat('f-measure: ')  
 cat(fmeasure \* 100)  
 cat('%')  
 cat('\n')  
}  
  
## Prediction Accuracy  
glm.predAcc <- mean(glm.pred == test$Churn1)  
glm.missClassError <- 1-glm.predAcc  
  
## Data is unbalanced - compute precision and accuracy  
glm.precision <- mean(test$Churn1[glm.pred == "1"] == "1") ## Precision - true + out of all predicted +  
glm.recall <- mean(glm.pred[test$Churn1 == "1"] == "1") ## Recall - true + out of all actual +  
  
measurePrecisionRecall(as.numeric(glm.pred), as.numeric(test$Churn1))

## precision: 64.21569%  
## recall: 46.78571%  
## f-measure: 54.13223%

caret::confusionMatrix(glm.pred, test$Churn1, positive = "1")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1402 298  
## 1 146 262  
##   
## Accuracy : 0.7894   
## 95% CI : (0.7713, 0.8066)  
## No Information Rate : 0.7343   
## P-Value [Acc > NIR] : 2.734e-09   
##   
## Kappa : 0.409   
## Mcnemar's Test P-Value : 7.714e-13   
##   
## Sensitivity : 0.4679   
## Specificity : 0.9057   
## Pos Pred Value : 0.6422   
## Neg Pred Value : 0.8247   
## Prevalence : 0.2657   
## Detection Rate : 0.1243   
## Detection Prevalence : 0.1935   
## Balanced Accuracy : 0.6868   
##   
## 'Positive' Class : 1   
##

## Prediction accuracy is around 79%. Precision is 64,2%.  
  
##========================  
## 2. Decision Tree  
##========================  
  
## partikit for more intuitive plot - overrides party package  
library(partykit)

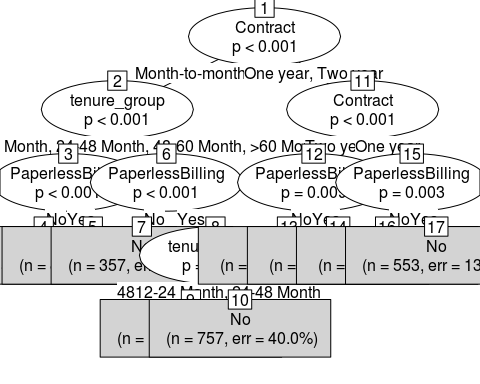
##   
## Attaching package: 'partykit'

## The following objects are masked from 'package:party':  
##   
## cforest, ctree, ctree\_control, edge\_simple, mob, mob\_control,  
## node\_barplot, node\_bivplot, node\_boxplot, node\_inner,  
## node\_surv, node\_terminal

dt.fit1 <- ctree(Churn ~ Contract + tenure\_group + PaperlessBilling, data = train)  
dt.fit1

##   
## Model formula:  
## Churn ~ Contract + tenure\_group + PaperlessBilling  
##   
## Fitted party:  
## [1] root  
## | [2] Contract in Month-to-month  
## | | [3] tenure\_group in 0-12 Month  
## | | | [4] PaperlessBilling in No: No (n = 530, err = 39.6%)  
## | | | [5] PaperlessBilling in Yes: Yes (n = 865, err = 40.0%)  
## | | [6] tenure\_group in 12-24 Month, 24-48 Month, 48-60 Month, >60 Month  
## | | | [7] PaperlessBilling in No: No (n = 357, err = 23.2%)  
## | | | [8] PaperlessBilling in Yes  
## | | | | [9] tenure\_group in 48-60 Month, >60 Month: No (n = 197, err = 27.4%)  
## | | | | [10] tenure\_group in 12-24 Month, 24-48 Month: No (n = 757, err = 40.0%)  
## | [11] Contract in One year, Two year  
## | | [12] Contract in Two year  
## | | | [13] PaperlessBilling in No: No (n = 623, err = 1.6%)  
## | | | [14] PaperlessBilling in Yes: No (n = 571, err = 4.6%)  
## | | [15] Contract in One year  
## | | | [16] PaperlessBilling in No: No (n = 471, err = 6.8%)  
## | | | [17] PaperlessBilling in Yes: No (n = 553, err = 13.0%)  
##   
## Number of inner nodes: 8  
## Number of terminal nodes: 9

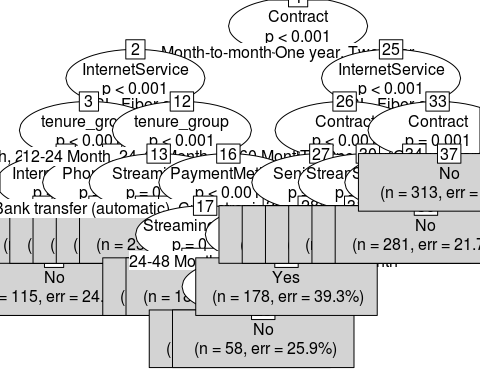
# plot(as.simpleparty(dt.fit))  
plot(dt.fit1, type = "simple")



dt.pred1 <- predict(dt.fit1, test)  
dt.predAcc1 <- mean(dt.pred1 == test$Churn)  
## Prediction accuracy of 76.3%  
  
dt.fit2 <- ctree(Churn ~ ., data = train)  
dt.fit2

##   
## Model formula:  
## Churn ~ gender + SeniorCitizen + Partner + Dependents + PhoneService +   
## MultipleLines + InternetService + OnlineSecurity + OnlineBackup +   
## DeviceProtection + TechSupport + StreamingTV + StreamingMovies +   
## Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +   
## tenure\_group  
##   
## Fitted party:  
## [1] root  
## | [2] Contract in Month-to-month  
## | | [3] InternetService in DSL, No  
## | | | [4] tenure\_group in 0-12 Month  
## | | | | [5] InternetService in DSL  
## | | | | | [6] MonthlyCharges <= 55.4: No (n = 361, err = 47.9%)  
## | | | | | [7] MonthlyCharges > 55.4: No (n = 115, err = 24.3%)  
## | | | | [8] InternetService in No: No (n = 264, err = 23.1%)  
## | | | [9] tenure\_group in 12-24 Month, 24-48 Month, 48-60 Month, >60 Month  
## | | | | [10] PhoneService in No: No (n = 130, err = 26.9%)  
## | | | | [11] PhoneService in Yes: No (n = 323, err = 13.0%)  
## | | [12] InternetService in Fiber optic  
## | | | [13] tenure\_group in 0-12 Month  
## | | | | [14] StreamingTV in No: Yes (n = 416, err = 33.2%)  
## | | | | [15] StreamingTV in Yes: Yes (n = 239, err = 20.9%)  
## | | | [16] tenure\_group in 12-24 Month, 24-48 Month, 48-60 Month, >60 Month  
## | | | | [17] PaymentMethod in Bank transfer (automatic), Credit card (automatic), Mailed check  
## | | | | | [18] StreamingMovies in No: No (n = 191, err = 23.0%)  
## | | | | | [19] StreamingMovies in Yes: No (n = 183, err = 41.0%)  
## | | | | [20] PaymentMethod in Electronic check  
## | | | | | [21] tenure\_group in 24-48 Month, 48-60 Month, >60 Month  
## | | | | | | [22] OnlineSecurity in No: No (n = 248, err = 48.8%)  
## | | | | | | [23] OnlineSecurity in Yes: No (n = 58, err = 25.9%)  
## | | | | | [24] tenure\_group in 12-24 Month: Yes (n = 178, err = 39.3%)  
## | [25] Contract in One year, Two year  
## | | [26] InternetService in DSL, No  
## | | | [27] Contract in Two year  
## | | | | [28] SeniorCitizen in No: No (n = 822, err = 1.1%)  
## | | | | [29] SeniorCitizen in Yes: No (n = 59, err = 6.8%)  
## | | | [30] Contract in One year  
## | | | | [31] StreamingMovies in No: No (n = 474, err = 4.0%)  
## | | | | [32] StreamingMovies in Yes: No (n = 175, err = 10.9%)  
## | | [33] InternetService in Fiber optic  
## | | | [34] Contract in One year  
## | | | | [35] StreamingMovies in No: No (n = 94, err = 5.3%)  
## | | | | [36] StreamingMovies in Yes: No (n = 281, err = 21.7%)  
## | | | [37] Contract in Two year: No (n = 313, err = 7.3%)  
##   
## Number of inner nodes: 18  
## Number of terminal nodes: 19

plot(dt.fit2, type = "simple")



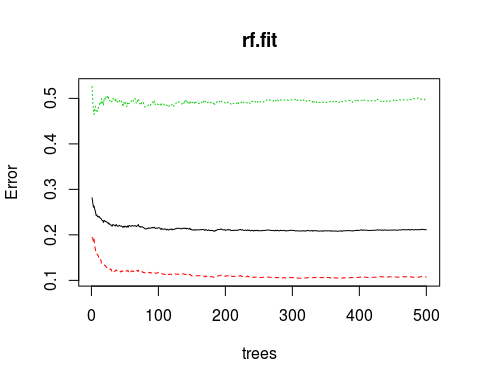
dt.pred2 <- predict(dt.fit2, test)  
dt.predAcc2 <- mean(dt.pred2 == test$Churn)  
  
## Prediction accuracy of 78%. Improvedment from the earlier model  
  
##============================  
## 3. Random Forest  
##=============================  
  
rf.fit <- randomForest(Churn ~ . , data = train)  
print(rf.fit)

##   
## Call:  
## randomForest(formula = Churn ~ ., data = train)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 4  
##   
## OOB estimate of error rate: 21.1%  
## Confusion matrix:  
## No Yes class.error  
## No 3227 388 0.1073306  
## Yes 651 658 0.4973262

## Out-of-bag error is 20.71%  
  
## Use this model to predict churn  
rf.pred <- predict(rf.fit, test)  
caret::confusionMatrix(rf.pred, test$Churn, positive = "Yes")

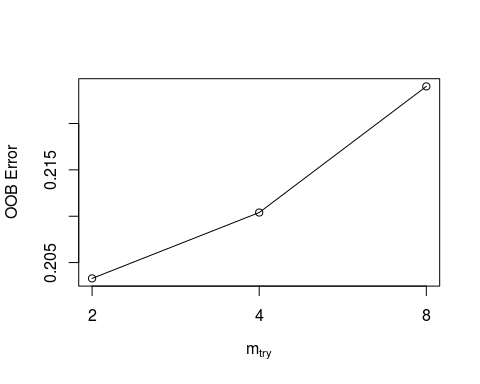
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1372 281  
## Yes 176 279  
##   
## Accuracy : 0.7832   
## 95% CI : (0.765, 0.8006)  
## No Information Rate : 0.7343   
## P-Value [Acc > NIR] : 1.237e-07   
##   
## Kappa : 0.409   
## Mcnemar's Test P-Value : 1.145e-06   
##   
## Sensitivity : 0.4982   
## Specificity : 0.8863   
## Pos Pred Value : 0.6132   
## Neg Pred Value : 0.8300   
## Prevalence : 0.2657   
## Detection Rate : 0.1324   
## Detection Prevalence : 0.2158   
## Balanced Accuracy : 0.6923   
##   
## 'Positive' Class : Yes   
##

## Prediction accuracy is 78.5%. Recall is 49.5%. i.e. of all actual churn cases in the test data only about 50% is predicted correctly.  
  
## Plot the error rate  
plot(rf.fit) ## plots the error rate vs. the number of trees



## As the number of trees increase, the OOB error decreases and then it stays constant.  
  
## Tune random forest model  
## Search for optimal value of mtry for randomForest()  
rf.tune <- tuneRF(train[, -18], train[, 18], stepFactor = 0.5, plot = TRUE, ntreeTry = 200, trace = TRUE, improve = 0.05)

## mtry = 4 OOB error = 21.04%   
## Searching left ...  
## mtry = 8 OOB error = 22.4%   
## -0.06467181 0.05   
## Searching right ...  
## mtry = 2 OOB error = 20.33%   
## 0.03378378 0.05



## Fit the model after tuning - take cue from the previous two parameters  
rf.fit.new <- randomForest(Churn ~ ., data = train, ntree = 200, mtry = 2, importance = T, proximity = T)  
print(rf.fit.new)

##   
## Call:  
## randomForest(formula = Churn ~ ., data = train, ntree = 200, mtry = 2, importance = T, proximity = T)   
## Type of random forest: classification  
## Number of trees: 200  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 20.49%  
## Confusion matrix:  
## No Yes class.error  
## No 3283 332 0.09183956  
## Yes 677 632 0.51718869

## Error rate is slightly lower.  
## Precdiction and Confusion matrix after tuning  
rf.pred.new <- predict(rf.fit.new, test)  
caret::confusionMatrix(rf.pred.new, test$Churn, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 1407 311  
## Yes 141 249  
##   
## Accuracy : 0.7856   
## 95% CI : (0.7674, 0.8029)  
## No Information Rate : 0.7343   
## P-Value [Acc > NIR] : 3.016e-08   
##   
## Kappa : 0.3915   
## Mcnemar's Test P-Value : 1.879e-15   
##   
## Sensitivity : 0.4446   
## Specificity : 0.9089   
## Pos Pred Value : 0.6385   
## Neg Pred Value : 0.8190   
## Prevalence : 0.2657   
## Detection Rate : 0.1181   
## Detection Prevalence : 0.1850   
## Balanced Accuracy : 0.6768   
##   
## 'Positive' Class : Yes   
##

## Accuracy is still around 78.5%  
  
## variable importance plot  
varImpPlot(rf.fit.new, sort = T, main = "Variable Importance")

