Telco Churn Analysis

Contents

[Introduction 3](#_Toc45188903)

[The Dataset 4](#_Toc45188904)

[Data Exploration 5](#_Toc45188905)

[Customer Characteristics 5](#_Toc45188906)

[Customer’s Services 7](#_Toc45188907)

[Glossary 10](#_Toc45188908)

[Code 11](#_Toc45188909)

# Introduction

Based on data provided by [Kaggle.com](https://www.kaggle.com/), predictive analysis was performed on the information provided by Telco and their customer churn.

In this project, the dataset was used as a means of continuation of my development of technical skills and use of predictive modeling to determine the overall outcome of a customer for Telco. This analysis contains methods of data cleaning, data engineering, handling imbalanced data for imbalanced dependent variables for classification modeling, and model selection. This paper will also contain elementary explanations and reasonings for the methods being used in order to practice the statistical, mathematical, and logical concepts within this field and to increase my overall capacity to create, develop, and provide efficacious models promptly; in addition, this will be used as a possible reference for future work.

Moreover, the topic of this analysis is ***customer churn***, a rudimentary principle within marketing describing the loss of customers. In relation to businesses, churn can have a large impact on the success of a business, because: it can decrease ***customer acquisition costs (CAC)*** by increasing the likelihood for a product to spread word of mouth reducing the direct costs of marketing; proper understanding of a company’s churn rate and the factors that contribute to customer churn can increase customer retention if those problems (problems which include but is not limited to wrong target audience, limitation of product’s desired effects, and poor customer experiences with the company); and it can increase the lifetime value of a customer (LTV) due to the customers loyalty and customer willingness to spend resources on company products. These factors both directly and indirectly contributes to a company’s success.

# The Dataset

This dataset contains 7043 observations with 20 independent variables and one dependent variable labeled as such:

|  |  |
| --- | --- |
| Variable Name: | Description: |
| customerID | A variable with a unique string made in order to identify individual customers |
| gender | Denoting whether the customer was either male of female  (Male/Female) |
| SeniorCitizen | Denoting whether the customer was a senior citizen or not (Yes/No) |
| Partner | Denoting whether the customer had a partner or not (Yes/No) |
| Dependents | Denoting whether the customer had any children or individuals who were relies on customer (Yes/No) |
| Tenure | A numerical variable which quantifies the length of period the customer has been with Telco |
| PhoneService | Denoting whether the customer used Telco phone services (Yes/No) |
| MultipleLines | Denoting whether the customer had one or more phone lines (Yes/No) |
| Internet Service | Denoting whether the customer was a part of Telco internet services (Yes/No) |
| Online Security | Denoting whether the Telco provided the customer’s internet security services (Yes/No) |
| Online Backup | Denoting whether the customer used Telco’s online backup services to secure their data from hardware malfunctions (Yes/No) |
| Device Protection | Denoting whether the device was insured by Telco (Yes/No) |
| TechSupport | N/A |
| StreamingTV | Denoting whether the customer streamed TV (Yes/No) |
| StreamingMovies | Denoting whether the customer streamed movies (Yes/No) |
| Contract | The length/type of contract (Month-to-Month/One-Year/Two-Year) |
| PaperlessBilling | Whether the customer was billed electronically (Yes/No) |
| Payment Method | The method in which the customer would pay for the services (Electronic check/ Mailed check/Bank transfer/Credit card) |
| MonthlyCharges | A numerical variable quantifying the monthly billing amount |
| TotalCharges | A numerical variable quantifying the total billing amount |
| Churn | The dependent variable indicating whether the customer discontinued services or not |

# Data Exploration

## Customer Characteristics

There are six charts being presented containing data about the customer’s characteristics including the methods in which the customer has decided to be billed. Each chart represents one factor of the customer characteristics which was been sub-grouped by whether that customer has churned or not.

|  |  |
| --- | --- |
| Figure 1 | Figure 2 |

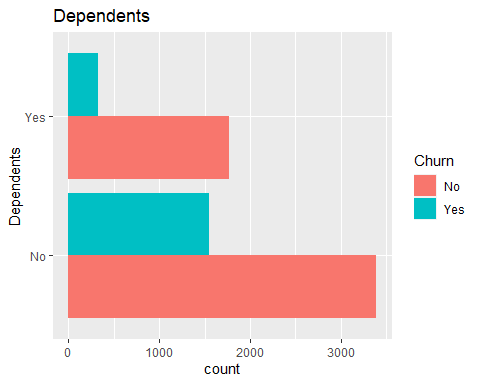


Figure 3

In regards to Figure 1 and Figure 2, not much information can be deciphered from this, as it appears that the proportions of customer churn between genders are approximately equal from a visual perspective and a mathematical approach must be taken to determine its statistical relevance. In addition, the same can be said for partner; however, it does look like those who do not have a partner may have a higher churn rate.

In Figure 3, there does appear to be a difference in the proportion of churn for individuals with dependents and those without dependents. This is probably likely due to the need for stability, meaning: those with dependents such as a child do not have much flexibility for changing services; in addition, those with dependents are less likely to move and are in a more stable period in their lives compared to those without independents; whereas, a younger person is more likely to move and/or is willing to have periods without services when switching to a competitor.

|  |
| --- |
| Figure 4 |

Through the visual analysis of Figure 4, it can be seen that senior citizens have a higher proportion of individuals churn when comparing against individuals who are not seniors. This can be rationally explained by the fact that seniors may be limited by their life expectancy and life events. For instance, a senior would have a higher chance of dying or be moved into a nursing facility or retirement home and in either situation the customer would likely terminate their services.

|  |  |
| --- | --- |
| Figure 5 | Figure 6 |

Based Figure 5 and Figure 6, it is difficult to determine a narrative for both of the charts presented. However, it appears that people who have paperless billing were experience a higher churn rate and individuals who paid through electronic check had a higher churn rate compared against the other payment methods.

## Customer’s Data

|  |  |
| --- | --- |
| Figure 7 | |
| Figure 8 | Figure 9 |

The three charts above, represents the services a customer has in relation to internet services. These variables originally had three factors, however, individuals who did not have internet service were added to the level “No”: meaning that some of the individuals in this level either do not have internet or they have internet and do not have the service.

Collectively,

|  |  |
| --- | --- |
| Figure 10 | Figure 11 |

Based on the charts presented above (Figure 7 and Figure 8), there is not a clear difference the proportions between the customer outcomes based on the services presented in the graphics.

|  |
| --- |
| Figure 12 |
| Figure 13 |

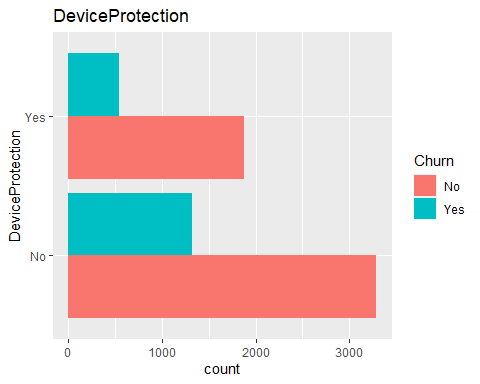


Figure 14

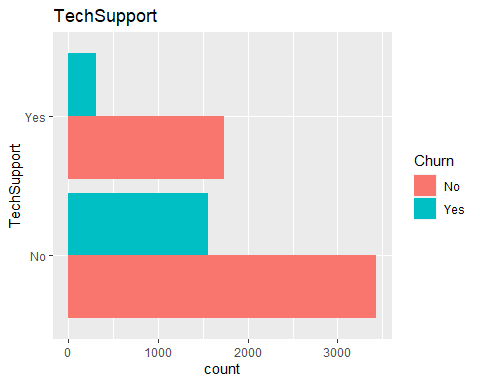
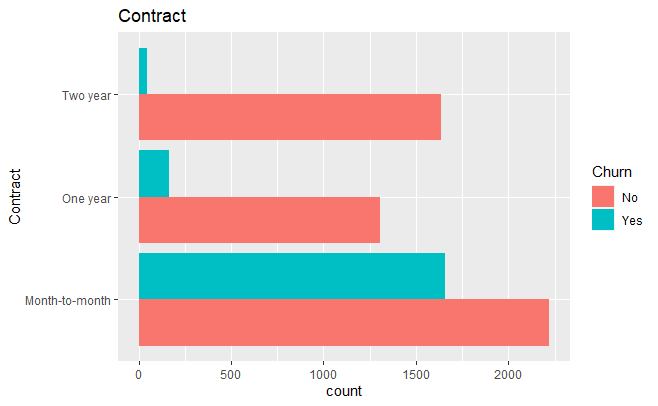


Figure 15



# Model Selection

# Model Results

## Boosted Logistic Regression:

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 843 138

Yes 182 244

Accuracy : 0.7726

95% CI : (0.7498, 0.7942)

No Information Rate : 0.7285

P-Value [Acc > NIR] : 8.846e-05

Kappa : 0.4451

Mcnemar's Test P-Value : 0.01623

Sensitivity : 0.8224

Specificity : 0.6387

Pos Pred Value : 0.8593

Neg Pred Value : 0.5728

Prevalence : 0.7285

Detection Rate : 0.5991

Detection Prevalence : 0.6972

Balanced Accuracy : 0.7306

'Positive' Class : No

## Classification Tree (Bagged):

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 936 211

Yes 89 171

Accuracy : 0.7868

95% CI : (0.7644, 0.8079)

No Information Rate : 0.7285

P-Value [Acc > NIR] : 2.806e-07

Kappa : 0.401

Mcnemar's Test P-Value : 2.830e-12

Sensitivity : 0.9132

Specificity : 0.4476

Pos Pred Value : 0.8160

Neg Pred Value : 0.6577

Prevalence : 0.7285

Detection Rate : 0.6652

Detection Prevalence : 0.8152

Balanced Accuracy : 0.6804

'Positive' Class : No

## Classification Tree:

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 970 242

Yes 55 140

Accuracy : 0.7889

95% CI : (0.7666, 0.81)

No Information Rate : 0.7285

P-Value [Acc > NIR] : 1.03e-07

Kappa : 0.3696

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9463

Specificity : 0.3665

Pos Pred Value : 0.8003

Neg Pred Value : 0.7179

Prevalence : 0.7285

Detection Rate : 0.6894

Detection Prevalence : 0.8614

Balanced Accuracy : 0.6564

'Positive' Class : No

## Random Forrest:

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 901 145

Yes 124 237

Accuracy : 0.8088

95% CI : (0.7873, 0.8291)

No Information Rate : 0.7285

P-Value [Acc > NIR] : 1.406e-12

Kappa : 0.5082

Mcnemar's Test P-Value : 0.2227

Sensitivity : 0.8790

Specificity : 0.6204

Pos Pred Value : 0.8614

Neg Pred Value : 0.6565

Prevalence : 0.7285

Detection Rate : 0.6404

Detection Prevalence : 0.7434

Balanced Accuracy : 0.7497

'Positive' Class : No

## Weighted Gradient Boosting:

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 914 152

Yes 111 230

Accuracy : 0.8131

95% CI : (0.7917, 0.8331)

No Information Rate : 0.7285

P-Value [Acc > NIR] : 8.124e-14

Kappa : 0.511

Mcnemar's Test P-Value : 0.01364

Sensitivity : 0.8917

Specificity : 0.6021

Pos Pred Value : 0.8574

Neg Pred Value : 0.6745

Prevalence : 0.7285

Detection Rate : 0.6496

Detection Prevalence : 0.7576

Balanced Accuracy : 0.7469

'Positive' Class : No

## Linear Discriminant Analysis

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 763 77

Yes 262 305

Accuracy : 0.7591

95% CI : (0.7358, 0.7812)

No Information Rate : 0.7285

P-Value [Acc > NIR] : 0.00502

Kappa : 0.4712

Mcnemar's Test P-Value : < 2e-16

Sensitivity : 0.7444

Specificity : 0.7984

Pos Pred Value : 0.9083

Neg Pred Value : 0.5379

Prevalence : 0.7285

Detection Rate : 0.5423

Detection Prevalence : 0.5970

Balanced Accuracy : 0.7714

'Positive' Class : No

## Linear Discriminant Analysis (Under sampling)

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 2926 98

Yes 742 276

Accuracy : 0.7922

95% CI : (0.7793, 0.8046)

No Information Rate : 0.9075

P-Value [Acc > NIR] : 1

Kappa : 0.3021

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.7977

Specificity : 0.7380

Pos Pred Value : 0.9676

Neg Pred Value : 0.2711

Prevalence : 0.9075

Detection Rate : 0.7239

Detection Prevalence : 0.7481

Balanced Accuracy : 0.7678

'Positive' Class : No

## GBM (Under sampling)

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 936 122

Yes 164 252

Accuracy : 0.806

95% CI : (0.7848, 0.8259)

No Information Rate : 0.7463

P-Value [Acc > NIR] : 3.6e-08

Kappa : 0.506

Mcnemar's Test P-Value : 0.01533

Sensitivity : 0.8509

Specificity : 0.6738

Pos Pred Value : 0.8847

Neg Pred Value : 0.6058

Prevalence : 0.7463

Detection Rate : 0.6350

Detection Prevalence : 0.7178

Balanced Accuracy : 0.7624

'Positive' Class : No

## Random Forrest (Under sampling)

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 2757 50

Yes 911 324

Accuracy : 0.7622

95% CI : (0.7488, 0.7753)

No Information Rate : 0.9075

P-Value [Acc > NIR] : 1

Kappa : 0.3039

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.7516

Specificity : 0.8663

Pos Pred Value : 0.9822

Neg Pred Value : 0.2623

Prevalence : 0.9075

Detection Rate : 0.6821

Detection Prevalence : 0.6945

Balanced Accuracy : 0.8090

'Positive' Class : No

# Glossary

Customer Churn: The loss of customers, commonly referenced companies who offer subscription services. 3

# Code

# Churn Analysis:

#

#

#

#

#

#

### Packages

library(ROSE)

library(caret)

library(e1071)

library(lattice)

library(pROC)

library(ggplot2)

library(plotROC)

library(tidyverse)

library(dplyr)

library(ROCR)

library(rpart)

library(rattle)

library(rpart.plot)

library(caTools)

library(pROC)

### Preliminary Functions

undersample\_ds <- function(x, classCol, nsamples\_class){

for (i in 1:length(unique(x[, classCol]))){

class.i <- unique(x[, classCol])[i]

if((sum(x[, classCol] == class.i) - nsamples\_class) != 0){

x <- x[-sample(which(x[, classCol] == class.i),

sum(x[, classCol] == class.i) - nsamples\_class), ]

}

}

return(x)

}

### Reading in Data

churn <- na.omit(read.csv("C:/Users/jeanp/OneDrive/Documents/GitHub/Telco\_Churn\_Analysis/WA\_Fn-UseC\_-Telco-Customer-Churn.csv"))

summary(churn)

str(churn)

### Recoding Data to reduce number of redundent factors within each Variable

# For this sitation cols 8, 10:15 have three factors Yes, no, and "No phone service" or "No Internet Service", respectively.

churn[c(8,10:15)] <- ifelse(churn[c(8,10:15)] == "Yes", "Yes" , "No")

churn[sapply(churn, is.character)] <- lapply(churn[sapply(churn, is.character)], as.factor)

# Col 3 was originally an indicator, however, to match the respective data it was returned to a str.

churn[,3] <- as.factor(ifelse(churn[,3] == 1, "Yes" , "No"))

# Ensuring that the base group for churn is "No" to predict the individuals who churn

churn$Churn <- relevel(churn$Churn, ref = "No")

#########################################################################################

### Data Exploration:

for(i in 2:20){

temp0 <- names(churn)

if(is.factor(churn[,i]) == TRUE){

gg\_plot <- ggplot(churn, aes(churn[,i], fill = Churn)) +

geom\_bar(position = position\_dodge()) +

ggtitle(temp0[i]) +

xlab(temp0[i]) +

coord\_flip()

print(gg\_plot)

}

}

# Creating a Bar Graph for the binomial factor "Churn"

plot(churn$Churn, main = "Churn")

table(churn$Churn)

t\_p <- data.frame(Response = c("No", "Yes"), Count = c(5174, 1869))

pie(x = t\_p$Count, labels = t\_p$Response, main = "Churn")

# Based on the Bar Chart given above, the data is not balanced and needs to be addressed or the data may face baises:

table(churn$Churn)

#########################################################################################

# Univariate Visualization

test <- glm(Churn ~ MonthlyCharges, data = churn, family = binomial(link = "logit"))

prob <- predict.glm(test, type = "response")

plot(churn$MonthlyCharges, ifelse(churn$Churn == "Yes", 1, 0))

curve(predict.glm(test,data.frame(MonthlyCharges=x), type = "response"), add = TRUE)

#########################################################################################

# Model Building (Backward Selection)

Model0 <- glm(Churn ~ 1, data = churn, family = binomial(link = logit))

CompleteModel <- glm(Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure + PhoneService + MultipleLines + InternetService + OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + StreamingMovies + StreamingTV + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges, data = churn, family = binomial(link = "logit"))

ReducedModel <- step(Model0, direction = 'both', scope = formula(CompleteModel))

summary(ReducedModel)

# Testing Model with ANOVA against the saturated model and the null model against the reduced model

Model\_test <- anova(ReducedModel, CompleteModel)

Model\_test0 <- anova(Model0, CompleteModel)

pchisq(Model\_test$Deviance, Model\_test$Df, lower.tail=FALSE)

pchisq(Model\_test0$Deviance, Model\_test0$Df, lower.tail=FALSE)

# Looking at combining Streaming

summary(Model\_streamingTV <- glm(Churn ~ StreamingTV, data = churn, family = binomial))

summary(Model\_streamingMovies <- glm(Churn ~ StreamingMovies, data = churn, family = binomial))

summary(Model\_streaming <- glm(Churn ~ StreamingMovies + StreamingTV, data = churn, family = binomial))

churn$streaming <- ifelse(churn$StreamingTV == "Yes"| churn$StreamingMovies == "Yes", "Yes", "No")

#Reconsolidating data based on results of the binomial model

churn$PaymentMethod2 <- as.factor(ifelse(data$PaymentMethod == "Electronic check", "Electronic check", "Other"))

ReducedModel2 <- glm(Churn ~ SeniorCitizen + streaming + Dependents + tenure + PhoneService + MultipleLines + OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + PaperlessBilling + PaymentMethod2 + MonthlyCharges, data = churn, family = binomial(link = "logit"))

summary(ReducedModel2)

# Model Building 2 (Backward Selection)

Model0 <- glm(Churn ~ 1, data = churn, family = binomial(link = logit))

CompleteModel <- glm(Churn ~ gender + SeniorCitizen + streaming + Partner + Dependents + tenure + PhoneService + MultipleLines + InternetService + OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + Contract + PaperlessBilling + PaymentMethod2 + MonthlyCharges, data = churn, family = binomial(link = "logit"))

ReducedModel <- step(Model0, direction = 'both', scope = formula(CompleteModel))

summary(ReducedModel)

#########################################################################################

##################### Predictive Modeling #####################

#########################################################################################

# Setting seed for reproducibility

set.seed(1)

train\_id <- sample(1:nrow(churn),nrow(churn)\*.8 ,replace=FALSE)

traind <- churn[train\_id, ]

testd <- churn[-train\_id, ]

plot(traind$Churn)

# creating model weights to deal with imbalanced response data

model\_weights <- ifelse(traind$Churn == "Yes", (1/table(traind$Churn)[1]) \* 0.5,(1/table(traind$Churn)[2]) \* 0.5)

# Control for Caret function (cross validation)

ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5, summaryFunction = twoClassSummary, classProbs = TRUE, seeds = )

#########################################################################################

# Logistic Regression (Boosted)

Boostlogit\_model <- train(Churn ~ Contract + InternetService + tenure + PaymentMethod +

PaperlessBilling + OnlineSecurity + TechSupport +

streaming + PhoneService + MultipleLines + SeniorCitizen +

Dependents + OnlineBackup,

data = traind,

method = "LogitBoost",

metric = "ROC",

wt = model\_weights,

trControl = ctrl)

prob <- predict.train(Baggedlogit\_model, newdata = testd,type = "prob")

pred <- ifelse(prob[,2] > .7, "Yes", "No")

table(pred, testd$Churn)

pred\_f <- as.factor(pred)

confusionMatrix(pred\_f, testd$Churn)

# Recieving Operating Characteristics for CART (Bagged)

pred = prediction(prob[,2], testd$Churn)

perf = performance(pred, "acc")

plot(perf)

roc = performance(pred,"tpr","fpr")

plot(roc, colorize = T, lwd = 2)

abline(a = 0, b = 1)

#########################################################################################

# Classification Tree (Bagged)

BaggedCART\_model <- train(Churn ~ Contract + tenure + PaymentMethod2 + PaperlessBilling +

OnlineSecurity + streaming + TechSupport + PhoneService +

MonthlyCharges + OnlineBackup + SeniorCitizen + DeviceProtection +

Dependents + MultipleLines,

data = traind,

method = "treebag",

metric = "ROC",

wt = model\_weights,

verbose = FALSE,

trControl = ctrl)

prob <- predict.train(BaggedCART\_model, newdata = testd, type = "prob")

pred <- ifelse(prob[,2] > .58, "Yes", "No")

table(pred, testd$Churn)

pred\_f <- as.factor(pred)

confusionMatrix(pred\_f, testd$Churn)

# Recieving Operating Characteristics for CART (Bagged)

pred = prediction(prob[,2], testd$Churn)

perf = performance(pred, "acc")

plot(perf)

roc = performance(pred,"tpr","fpr")

plot(roc, colorize = T, lwd = 2)

abline(a = 0, b = 1)

#########################################################################################

# Classification Tree

TREE\_model <- train(Churn ~ Contract + tenure + PaymentMethod2 + PaperlessBilling +

OnlineSecurity + streaming + TechSupport + PhoneService +

MonthlyCharges + OnlineBackup + SeniorCitizen + DeviceProtection +

Dependents + MultipleLines,

data = traind,

method = "rpart",

metric = "ROC",

trControl = ctrl)

prob <- predict.train(TREE\_model, newdata = testd, type = "prob")

pred <- ifelse(prob[,2] > .5, "Yes", "No")

table(pred, testd$Churn)

pred\_f <- as.factor(pred)

confusionMatrix(pred\_f, testd$Churn)

# Recieving Operating Characteristics for CART (Bagged)

pred = prediction(prob[,2], testd$Churn)

perf = performance(pred, "acc")

plot(perf)

roc = performance(pred,"tpr","fpr")

plot(roc, colorize = T, lwd = 2)

abline(a = 0, b = 1)

#########################################################################################

# Random Forrest

rf\_model <- train(Churn ~ Contract + tenure + PaymentMethod2 + PaperlessBilling +

OnlineSecurity + streaming + TechSupport + PhoneService +

MonthlyCharges + OnlineBackup + SeniorCitizen + DeviceProtection +

Dependents + MultipleLines,

data = traind,

method = "rf",

metric = "ROC",

verbose = FALSE,

trControl = ctrl)

prob <- predict.train(rf\_model, newdata = testd, type = "prob")

pred <- ifelse(prob[,2] > .36, "Yes", "No")

table(pred, testd$Churn)

pred\_f <- as.factor(pred)

confusionMatrix(pred\_f, testd$Churn)

# Recieving Operating Characteristics for Random Forrest

pred = prediction(prob[,2], testd$Churn)

perf = performance(pred, "acc")

plot(perf)

roc = performance(pred,"tpr","fpr")

plot(roc, colorize = T, lwd = 2)

abline(a = 0, b = 1)

auc.tmp <- performance(pred,"auc")

auc <- as.numeric(auc.tmp@y.values)

#########################################################################################

# Weighted GBM

gbm\_model\_w <- train(Churn ~ Contract + tenure + PaymentMethod2 + PaperlessBilling +

OnlineSecurity + streaming + TechSupport + PhoneService +

MonthlyCharges + OnlineBackup + SeniorCitizen + DeviceProtection +

Dependents + MultipleLines,

data = traind,

method = "gbm",

verbose = FALSE,

weights = model\_weights,

metric = "ROC",

trControl = ctrl)

# Reciever Operating Characteristics for Gradient Boosted Model

prob <- predict.train(gbm\_model\_w, newdata = testd, type = "prob")

pred\_f <- ifelse(prob[,2] > .229, "Yes", "No")

table(pred\_f, testd$Churn)

pred\_f <- as.factor(pred\_f)

confusionMatrix(pred\_f, testd$Churn)

# Reciever Operating Characteristics for Weighted Gradient Boosted Model

pred = prediction(prob[,2], testd$Churn)

perf = performance(pred, "acc")

plot(perf)

roc = performance(pred,"tpr","fpr")

plot(roc, colorize = T, lwd = 2)

abline(a = 0, b = 1)

auc.tmp <- performance(pred,"auc")

auc <- as.numeric(auc.tmp@y.values)