Customer Churn Prediction

#Importing Libraries  
library(readxl)  
library(data.table)

## Warning: package 'data.table' was built under R version 4.0.3

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(caret)

## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.4

library(ROCR)

## Warning: package 'ROCR' was built under R version 4.0.4

#Setting the Working Directory and Importing the Dataset  
setwd("C:/Users/surya/Downloads")  
  
tc <- read\_excel("TelcoChurn.xlsx", sheet = "Data")  
names(tc) <- tolower(colnames(tc))  
attach(tc)  
  
#NA Values Column-Wise & Pre-Processing  
sapply(tc, 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

str(tc)

## tibble [7,043 x 21] (S3: tbl\_df/tbl/data.frame)  
## $ customerid : chr [1:7043] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...  
## $ gender : chr [1:7043] "Female" "Male" "Male" "Male" ...  
## $ seniorcitizen : num [1:7043] 0 0 0 0 0 0 0 0 0 0 ...  
## $ partner : chr [1:7043] "Yes" "No" "No" "No" ...  
## $ dependents : chr [1:7043] "No" "No" "No" "No" ...  
## $ tenure : num [1:7043] 1 34 2 45 2 8 22 10 28 62 ...  
## $ phoneservice : chr [1:7043] "No" "Yes" "Yes" "No" ...  
## $ multiplelines : chr [1:7043] "No phone service" "No" "No" "No phone service" ...  
## $ internetservice : chr [1:7043] "DSL" "DSL" "DSL" "DSL" ...  
## $ onlinesecurity : chr [1:7043] "No" "Yes" "Yes" "Yes" ...  
## $ onlinebackup : chr [1:7043] "Yes" "No" "Yes" "No" ...  
## $ deviceprotection: chr [1:7043] "No" "Yes" "No" "Yes" ...  
## $ techsupport : chr [1:7043] "No" "No" "No" "Yes" ...  
## $ streamingtv : chr [1:7043] "No" "No" "No" "No" ...  
## $ streamingmovies : chr [1:7043] "No" "No" "No" "No" ...  
## $ contract : chr [1:7043] "Month-to-month" "One year" "Month-to-month" "One year" ...  
## $ paperlessbilling: chr [1:7043] "Yes" "No" "Yes" "No" ...  
## $ paymentmethod : chr [1:7043] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...  
## $ monthlycharges : num [1:7043] 29.9 57 53.9 42.3 70.7 ...  
## $ totalcharges : num [1:7043] 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ churn : chr [1:7043] "No" "No" "Yes" "No" ...

colSums(is.na(tc))

## 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

tc <- tc[complete.cases(tc), ]  
str(tc)

## tibble [7,032 x 21] (S3: tbl\_df/tbl/data.frame)  
## $ customerid : chr [1:7032] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...  
## $ gender : chr [1:7032] "Female" "Male" "Male" "Male" ...  
## $ seniorcitizen : num [1:7032] 0 0 0 0 0 0 0 0 0 0 ...  
## $ partner : chr [1:7032] "Yes" "No" "No" "No" ...  
## $ dependents : chr [1:7032] "No" "No" "No" "No" ...  
## $ tenure : num [1:7032] 1 34 2 45 2 8 22 10 28 62 ...  
## $ phoneservice : chr [1:7032] "No" "Yes" "Yes" "No" ...  
## $ multiplelines : chr [1:7032] "No phone service" "No" "No" "No phone service" ...  
## $ internetservice : chr [1:7032] "DSL" "DSL" "DSL" "DSL" ...  
## $ onlinesecurity : chr [1:7032] "No" "Yes" "Yes" "Yes" ...  
## $ onlinebackup : chr [1:7032] "Yes" "No" "Yes" "No" ...  
## $ deviceprotection: chr [1:7032] "No" "Yes" "No" "Yes" ...  
## $ techsupport : chr [1:7032] "No" "No" "No" "Yes" ...  
## $ streamingtv : chr [1:7032] "No" "No" "No" "No" ...  
## $ streamingmovies : chr [1:7032] "No" "No" "No" "No" ...  
## $ contract : chr [1:7032] "Month-to-month" "One year" "Month-to-month" "One year" ...  
## $ paperlessbilling: chr [1:7032] "Yes" "No" "Yes" "No" ...  
## $ paymentmethod : chr [1:7032] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...  
## $ monthlycharges : num [1:7032] 29.9 57 53.9 42.3 70.7 ...  
## $ totalcharges : num [1:7032] 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ churn : chr [1:7032] "No" "No" "Yes" "No" ...

tc$customerid <- NULL  
tc$onlinebackup <- NULL  
tc$deviceprotection <-NULL  
tc$contract <- NULL  
tc$onlinesecurity <- NULL  
tc$streamingtv <- NULL  
tc$paperlessbilling <- NULL  
tc$paymentmethod <- NULL  
  
#Feature Engineering  
tc$churn <- ifelse(tc$churn == "Yes", 1, 0)  
tc$churn <- as.factor(tc$churn)  
  
phone\_service <- tc[tc$phoneservice == "Yes",]  
phone\_service <- phone\_service[, -c(6, 8:10)]  
  
internet\_service <- tc[tc$internetservice != "No",]  
internet\_service <- internet\_service[, -c(6:7)]  
  
both\_services <- tc[(tc$phoneservice == "Yes" & tc$internetservice != "No"),]  
both\_services <- both\_services[, -c(6)]  
  
#Exploratory Analysis  
table(tc$churn, tc$phoneservice)

##   
## No Yes  
## 0 510 4653  
## 1 170 1699

table(tc$churn, tc$internetservice)

##   
## DSL Fiber optic No  
## 0 1957 1799 1407  
## 1 459 1297 113

#Classification Models  
#Phone Service  
set.seed(1024)  
phone\_sample = floor(0.75\*nrow(phone\_service))  
index\_ps <- sample(seq\_len(nrow(phone\_service)), size=phone\_sample)  
train\_ps <- phone\_service[index\_ps,]  
test\_ps <- phone\_service[-index\_ps,]  
  
phone\_logit <- glm(churn ~ gender + seniorcitizen + partner + dependents + tenure + multiplelines +   
 monthlycharges + totalcharges, family = binomial (link = "logit"), data = train\_ps)  
str(test\_ps)

## tibble [1,588 x 9] (S3: tbl\_df/tbl/data.frame)  
## $ gender : chr [1:1588] "Male" "Male" "Male" "Male" ...  
## $ seniorcitizen : num [1:1588] 0 0 0 0 0 0 0 0 0 0 ...  
## $ partner : chr [1:1588] "Yes" "Yes" "No" "No" ...  
## $ dependents : chr [1:1588] "Yes" "No" "No" "No" ...  
## $ tenure : num [1:1588] 13 58 49 25 71 49 30 72 17 10 ...  
## $ multiplelines : chr [1:1588] "No" "Yes" "Yes" "No" ...  
## $ monthlycharges: num [1:1588] 50 100 104 106 107 ...  
## $ totalcharges : num [1:1588] 587 5681 5036 2686 7382 ...  
## $ churn : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 2 1 ...

test\_xps <- test\_ps[ , c(1:8)]  
predlogit\_phone <-predict(phone\_logit, newdata=test\_xps, type="response")  
predlogit\_phone <- ifelse(predlogit\_phone>0.5, 1, 0)  
  
#Internet Service  
set.seed(1024)  
internet\_sample = floor(0.75\*nrow(internet\_service))  
index\_is <- sample(seq\_len(nrow(internet\_service)), size=internet\_sample)  
train\_is <- internet\_service[index\_is,]  
test\_is <- internet\_service[-index\_is,]  
  
internet\_logit <- glm(churn ~ gender + seniorcitizen + partner + dependents + tenure + techsupport +   
 streamingmovies + monthlycharges + totalcharges, family = binomial (link = "logit"), data = train\_is)  
str(test\_is)

## tibble [1,378 x 11] (S3: tbl\_df/tbl/data.frame)  
## $ gender : chr [1:1378] "Female" "Male" "Male" "Male" ...  
## $ seniorcitizen : num [1:1378] 0 0 0 0 0 0 0 0 0 0 ...  
## $ partner : chr [1:1378] "No" "Yes" "Yes" "No" ...  
## $ dependents : chr [1:1378] "No" "Yes" "No" "No" ...  
## $ tenure : num [1:1378] 10 13 58 25 69 49 30 1 72 46 ...  
## $ internetservice: chr [1:1378] "DSL" "DSL" "Fiber optic" "Fiber optic" ...  
## $ techsupport : chr [1:1378] "No" "No" "No" "Yes" ...  
## $ streamingmovies: chr [1:1378] "No" "No" "Yes" "Yes" ...  
## $ monthlycharges : num [1:1378] 29.8 50 100.3 105.5 113.2 ...  
## $ totalcharges : num [1:1378] 302 587 5681 2686 7895 ...  
## $ churn : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...

test\_xis <- test\_is[ , c(1:10)]  
predlogit\_internet <-predict(internet\_logit, newdata=test\_xis, type="response")  
predlogit\_internet <- ifelse(predlogit\_internet>0.5, 1, 0)  
  
#Both Services  
set.seed(1024)  
both\_sample = floor(0.75\*nrow(both\_services))  
index\_bs <- sample(seq\_len(nrow(both\_services)), size=both\_sample)  
train\_bs <- both\_services[index\_bs,]  
test\_bs <- both\_services[-index\_bs,]  
  
both\_logit <- glm(churn ~ gender + seniorcitizen + partner + dependents + tenure + multiplelines +   
 techsupport + streamingmovies + monthlycharges + totalcharges, family = binomial (link = "logit"), data = train\_bs)  
str(test\_bs)

## tibble [1,208 x 12] (S3: tbl\_df/tbl/data.frame)  
## $ gender : chr [1:1208] "Male" "Male" "Female" "Female" ...  
## $ seniorcitizen : num [1:1208] 0 0 0 0 0 0 1 1 0 0 ...  
## $ partner : chr [1:1208] "No" "No" "Yes" "No" ...  
## $ dependents : chr [1:1208] "No" "No" "Yes" "No" ...  
## $ tenure : num [1:1208] 49 25 69 21 72 17 71 1 2 52 ...  
## $ multiplelines : chr [1:1208] "Yes" "No" "Yes" "No" ...  
## $ internetservice: chr [1:1208] "Fiber optic" "Fiber optic" "Fiber optic" "Fiber optic" ...  
## $ techsupport : chr [1:1208] "No" "Yes" "Yes" "No" ...  
## $ streamingmovies: chr [1:1208] "Yes" "Yes" "Yes" "Yes" ...  
## $ monthlycharges : num [1:1208] 103.7 105.5 113.2 90 90.2 ...  
## $ totalcharges : num [1:1208] 5036 2686 7895 1863 6369 ...  
## $ churn : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 1 ...

test\_xbs <- test\_bs[ , c(1:11)]  
predlogit\_both <-predict(both\_logit, newdata=test\_xbs, type="response")  
predlogit\_both <- ifelse(predlogit\_both>0.5, 1, 0)  
  
#Stargazer  
stargazer(phone\_logit, internet\_logit, both\_logit, type='text', single.row = TRUE)

##   
## =========================================================================  
## Dependent variable:   
## ------------------------------------------------------  
## churn   
## (1) (2) (3)   
## -------------------------------------------------------------------------  
## genderMale -0.004 (0.077) -0.040 (0.076) -0.056 (0.081)   
## seniorcitizen 0.405\*\*\* (0.099) 0.446\*\*\* (0.095) 0.339\*\*\* (0.100)   
## partnerYes 0.038 (0.092) 0.027 (0.090) -0.032 (0.096)   
## dependentsYes -0.372\*\*\* (0.105) -0.329\*\*\* (0.105) -0.162 (0.111)   
## tenure -0.084\*\*\* (0.008) -0.063\*\*\* (0.008) -0.074\*\*\* (0.010)  
## multiplelinesYes 0.373\*\*\* (0.092) 0.388\*\*\* (0.096)   
## techsupportYes -0.834\*\*\* (0.092) -0.708\*\*\* (0.099)  
## streamingmoviesYes 0.062 (0.093) -0.102 (0.105)   
## monthlycharges 0.029\*\*\* (0.002) 0.020\*\*\* (0.003) 0.027\*\*\* (0.004)   
## totalcharges 0.0003\*\*\* (0.0001) 0.0002\*\* (0.0001) 0.0003\*\* (0.0001)  
## Constant -1.655\*\*\* (0.163) -0.758\*\*\* (0.209) -1.384\*\*\* (0.292)  
## -------------------------------------------------------------------------  
## Observations 4,764 4,134 3,624   
## Log Likelihood -2,078.735 -2,060.649 -1,835.049   
## Akaike Inf. Crit. 4,175.470 4,141.299 3,692.099   
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

#Odds Ratio  
odd\_phone <- exp(phone\_logit$coef)  
odd\_internet <- exp(internet\_logit$coef)  
odd\_both <- exp(both\_logit$coef)  
  
e1 <- data.frame(odd\_phone)  
e2 <- data.frame(odd\_internet)  
e3 <- data.frame(odd\_both)  
  
#Probability  
prob\_phone <- exp(phone\_logit$coef)/(1 + exp(phone\_logit$coef))  
prob\_internet <- exp(internet\_logit$coef)/(1 + exp(internet\_logit$coef))  
prob\_both <- exp(both\_logit$coef)/(1 + exp(both\_logit$coef))  
  
p1 <- data.frame(prob\_phone)  
p2 <- data.frame(prob\_internet)  
p3 <- data.frame(prob\_both)  
  
#Summary of OR & Probability  
df1 <- data.frame(e1, p1)  
df2 <- data.frame(e2, p2)  
df3 <- data.frame(e3, p3)  
  
df1

## odd\_phone prob\_phone  
## (Intercept) 0.1911033 0.1604423  
## genderMale 0.9962687 0.4990654  
## seniorcitizen 1.4995681 0.5999309  
## partnerYes 1.0392097 0.5096140  
## dependentsYes 0.6891147 0.4079739  
## tenure 0.9191004 0.4789225  
## multiplelinesYes 1.4520795 0.5921829  
## monthlycharges 1.0298312 0.5073482  
## totalcharges 1.0002755 0.5000689

df2

## odd\_internet prob\_internet  
## (Intercept) 0.4685039 0.3190349  
## genderMale 0.9611909 0.4901057  
## seniorcitizen 1.5615696 0.6096144  
## partnerYes 1.0272919 0.5067311  
## dependentsYes 0.7197955 0.4185355  
## tenure 0.9386142 0.4841676  
## techsupportYes 0.4342869 0.3027894  
## streamingmoviesYes 1.0639819 0.5154996  
## monthlycharges 1.0204057 0.5050499  
## totalcharges 1.0001879 0.5000470

df3

## odd\_both prob\_both  
## (Intercept) 0.2506800 0.2004349  
## genderMale 0.9453833 0.4859625  
## seniorcitizen 1.4038160 0.5839948  
## partnerYes 0.9688884 0.4920992  
## dependentsYes 0.8504340 0.4595862  
## tenure 0.9287954 0.4815417  
## multiplelinesYes 1.4735288 0.5957193  
## techsupportYes 0.4923974 0.3299372  
## streamingmoviesYes 0.9030975 0.4745409  
## monthlycharges 1.0271267 0.5066909  
## totalcharges 1.0002556 0.5000639

#Model Fit & All Metrics  
#Phone Service  
table(test\_ps$churn, predlogit\_phone)

## predlogit\_phone  
## 0 1  
## 0 1047 124  
## 1 212 205

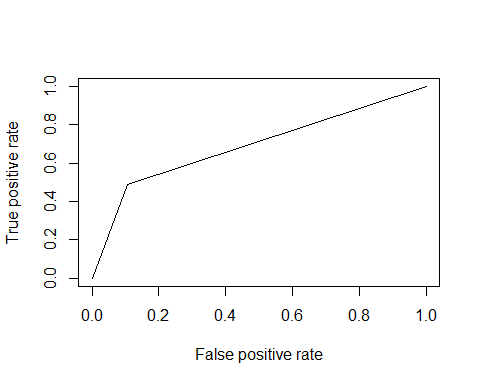
ClassificationError <- mean(predlogit\_phone != test\_ps$churn)  
print(paste("Accuracy = ", 1-ClassificationError))

## [1] "Accuracy = 0.788413098236776"

#Confusion Matrix, Predicted Accuracy, Recall, Precision, F1-Score & AUC  
cm <- confusionMatrix(as.factor(predlogit\_phone), reference = test\_ps$churn, mode = "everything")  
cm

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1047 212  
## 1 124 205  
##   
## Accuracy : 0.7884   
## 95% CI : (0.7675, 0.8083)  
## No Information Rate : 0.7374   
## P-Value [Acc > NIR] : 1.353e-06   
##   
## Kappa : 0.4138   
##   
## Mcnemar's Test P-Value : 2.072e-06   
##   
## Sensitivity : 0.8941   
## Specificity : 0.4916   
## Pos Pred Value : 0.8316   
## Neg Pred Value : 0.6231   
## Precision : 0.8316   
## Recall : 0.8941   
## F1 : 0.8617   
## Prevalence : 0.7374   
## Detection Rate : 0.6593   
## Detection Prevalence : 0.7928   
## Balanced Accuracy : 0.6929   
##   
## 'Positive' Class : 0   
##

#ROC  
ps <- prediction(predlogit\_phone, test\_ps$churn)  
ps\_perf <- performance(ps, measure="tpr", x.measure="fpr")  
plot(ps\_perf)



#AUC  
auc\_ps <- performance(ps, measure="auc")  
auc\_ps <- auc\_ps@y.values[[1]]  
auc\_ps

## [1] 0.6928572

#Internet Service  
table(test\_is$churn, predlogit\_internet)

## predlogit\_internet  
## 0 1  
## 0 816 141  
## 1 198 223

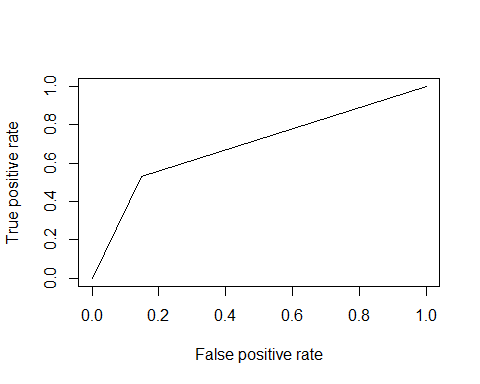
ClassificationError <- mean(predlogit\_internet != test\_is$churn)  
print(paste("Accuracy = ", 1-ClassificationError))

## [1] "Accuracy = 0.753991291727141"

#Confusion Matrix, Predicted Accuracy, Recall, Precision, F1-Score & AUC  
cm <- confusionMatrix(as.factor(predlogit\_internet), reference = test\_is$churn, mode = "everything")  
cm

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 816 198  
## 1 141 223  
##   
## Accuracy : 0.754   
## 95% CI : (0.7304, 0.7765)  
## No Information Rate : 0.6945   
## P-Value [Acc > NIR] : 5.874e-07   
##   
## Kappa : 0.3974   
##   
## Mcnemar's Test P-Value : 0.002354   
##   
## Sensitivity : 0.8527   
## Specificity : 0.5297   
## Pos Pred Value : 0.8047   
## Neg Pred Value : 0.6126   
## Precision : 0.8047   
## Recall : 0.8527   
## F1 : 0.8280   
## Prevalence : 0.6945   
## Detection Rate : 0.5922   
## Detection Prevalence : 0.7358   
## Balanced Accuracy : 0.6912   
##   
## 'Positive' Class : 0   
##

#ROC  
is <- prediction(predlogit\_internet, test\_is$churn)  
is\_perf <- performance(is, measure="tpr", x.measure="fpr")  
plot(is\_perf)



#AUC  
auc\_is <- performance(is, measure="auc")  
auc\_is <- auc\_is@y.values[[1]]  
auc\_is

## [1] 0.6911779

#Both Services  
table(test\_bs$churn, predlogit\_both)

## predlogit\_both  
## 0 1  
## 0 721 104  
## 1 186 197

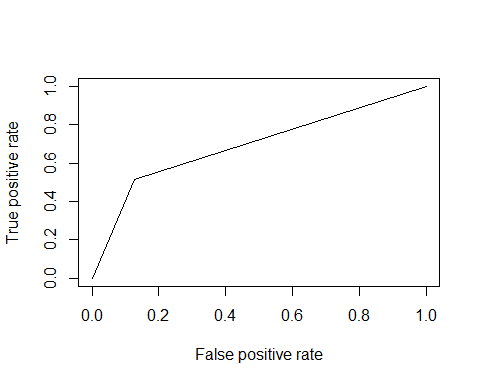
ClassificationError <- mean(predlogit\_both != test\_bs$churn)  
print(paste("Accuracy = ", 1-ClassificationError))

## [1] "Accuracy = 0.759933774834437"

#Confusion Matrix, Predicted Accuracy, Recall, Precision, F1-Score & AUC  
cm <- confusionMatrix(as.factor(predlogit\_both), reference = test\_bs$churn, mode = "everything")  
cm

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 721 186  
## 1 104 197  
##   
## Accuracy : 0.7599   
## 95% CI : (0.7348, 0.7838)  
## No Information Rate : 0.6829   
## P-Value [Acc > NIR] : 2.305e-09   
##   
## Kappa : 0.4119   
##   
## Mcnemar's Test P-Value : 1.970e-06   
##   
## Sensitivity : 0.8739   
## Specificity : 0.5144   
## Pos Pred Value : 0.7949   
## Neg Pred Value : 0.6545   
## Precision : 0.7949   
## Recall : 0.8739   
## F1 : 0.8326   
## Prevalence : 0.6829   
## Detection Rate : 0.5969   
## Detection Prevalence : 0.7508   
## Balanced Accuracy : 0.6941   
##   
## 'Positive' Class : 0   
##

#ROC  
bs <- prediction(predlogit\_both, test\_bs$churn)  
bs\_perf <- performance(bs, measure="tpr", x.measure="fpr")  
plot(bs\_perf)



#AUC  
auc\_bs <- performance(bs, measure="auc")  
auc\_bs <- auc\_bs@y.values[[1]]  
auc\_bs

## [1] 0.6941499