QWE

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

## Including Plots

library(readxl)  
QWE\_Excel\_1\_ <- read\_excel("QWE-Excel (1).xlsx",sheet = "Case Data")  
  
class(QWE\_Excel\_1\_)

## [1] "tbl\_df" "tbl" "data.frame"

qwe <- QWE\_Excel\_1\_   
  
library(mice)

## Loading required package: lattice

##   
## Attaching package: 'mice'

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(nnet)  
library(caret)

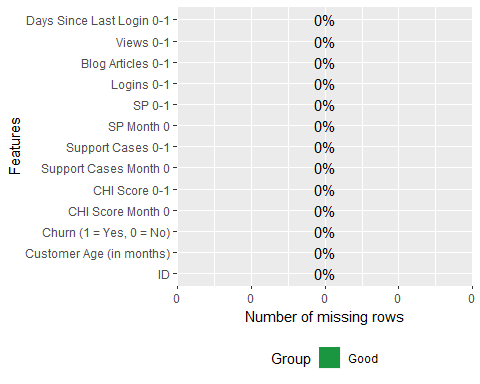
## Loading required package: ggplot2

library(DataExplorer)  
#install.packages('scales')  
library(scales)  
#install.packages('data.table')  
library(data.table)  
require(ggplot2)

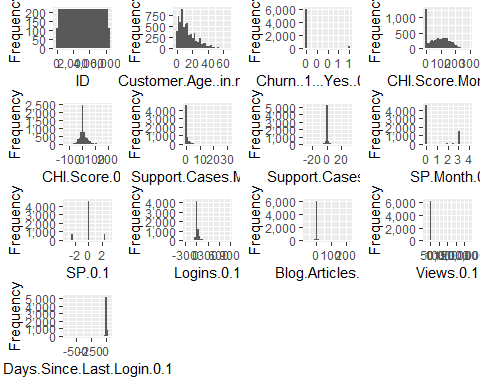
summary(QWE\_Excel\_1\_)

## ID Customer Age (in months) Churn (1 = Yes, 0 = No)  
## Min. : 1 Min. : 0.0 Min. :0.00000   
## 1st Qu.:1588 1st Qu.: 5.0 1st Qu.:0.00000   
## Median :3174 Median :11.0 Median :0.00000   
## Mean :3174 Mean :13.9 Mean :0.05089   
## 3rd Qu.:4760 3rd Qu.:20.0 3rd Qu.:0.00000   
## Max. :6347 Max. :67.0 Max. :1.00000   
## CHI Score Month 0 CHI Score 0-1 Support Cases Month 0  
## Min. : 0.00 Min. :-125.000 Min. : 0.0000   
## 1st Qu.: 24.50 1st Qu.: -8.000 1st Qu.: 0.0000   
## Median : 87.00 Median : 0.000 Median : 0.0000   
## Mean : 87.32 Mean : 5.059 Mean : 0.7063   
## 3rd Qu.:139.00 3rd Qu.: 15.000 3rd Qu.: 1.0000   
## Max. :298.00 Max. : 208.000 Max. :32.0000   
## Support Cases 0-1 SP Month 0 SP 0-1   
## Min. :-29.000000 Min. :0.0000 Min. :-4.00000   
## 1st Qu.: 0.000000 1st Qu.:0.0000 1st Qu.: 0.00000   
## Median : 0.000000 Median :0.0000 Median : 0.00000   
## Mean : -0.006932 Mean :0.8128 Mean : 0.03017   
## 3rd Qu.: 0.000000 3rd Qu.:2.6667 3rd Qu.: 0.00000   
## Max. : 31.000000 Max. :4.0000 Max. : 4.00000   
## Logins 0-1 Blog Articles 0-1 Views 0-1   
## Min. :-293.00 Min. :-75.0000 Min. :-28322.00   
## 1st Qu.: -1.00 1st Qu.: 0.0000 1st Qu.: -11.00   
## Median : 2.00 Median : 0.0000 Median : 0.00   
## Mean : 15.73 Mean : 0.1572 Mean : 96.31   
## 3rd Qu.: 23.00 3rd Qu.: 0.0000 3rd Qu.: 27.00   
## Max. : 865.00 Max. :217.0000 Max. :230414.00   
## Days Since Last Login 0-1  
## Min. :-648.000   
## 1st Qu.: 0.000   
## Median : 0.000   
## Mean : 1.765   
## 3rd Qu.: 3.000   
## Max. : 61.000

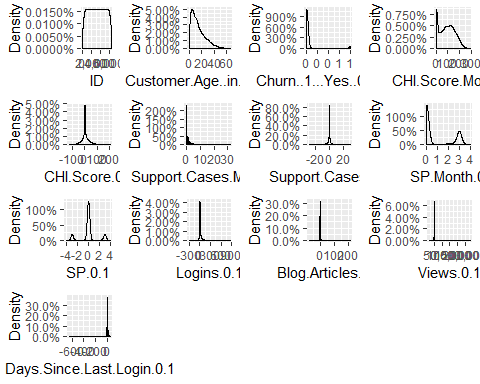
plot\_str(QWE\_Excel\_1\_) #we have all variable int tye  
plot\_missing(QWE\_Excel\_1\_) # 5 variables have Missing values



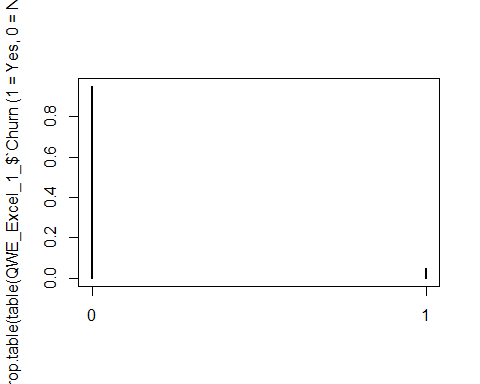
plot\_histogram(QWE\_Excel\_1\_) #to plot histogram of continuous variables



plot\_density(QWE\_Excel\_1\_) #to plot density plot of variable



plot(prop.table(table(QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)))



colnames(qwe)[which(names(qwe) == "CHI Score 0-1")] <- "CHI\_score"  
colnames(qwe)[which(names(qwe) == "Customer Age (in months)")] <- "customer\_age"  
colnames(qwe)[which(names(qwe) == "Churn (1 = Yes, 0 = No)")] <- "Churn"  
colnames(qwe)[which(names(qwe) == "CHI Score Month 0")] <- "churn\_score\_0"  
colnames(qwe)[which(names(qwe) == "Support Cases Month 0")] <- "Support\_Cases\_0"  
colnames(qwe)[which(names(qwe) == "CHI Score 0-1")] <- "CHI\_0\_1"  
colnames(qwe)[which(names(qwe) == "Support Cases 0-1")] <- "Support\_Cases\_0\_1"  
colnames(qwe)[which(names(qwe) == "SP Month 0")] <- "sp\_mon\_0"  
colnames(qwe)[which(names(qwe) == "Views 0-1")] <- "views\_0\_1"  
colnames(qwe)[which(names(qwe) == "SP 0-1")] <- "SP\_0\_1"  
colnames(qwe)[which(names(qwe) == "Logins 0-1")] <- "logins\_0\_1"  
colnames(qwe)[which(names(qwe) == "Blog Articles 0-1")] <- "blog\_articles\_0\_1"  
colnames(qwe)[which(names(qwe) == "Views 0-1")] <- "views\_0\_1"  
colnames(qwe)[which(names(qwe) == "Days Since Last Login 0-1")] <- "DaysSinceLast\_Login\_0\_1"

df <- qwe[,c(2,3,4,5,6,7,8,9,10,11,12,13)]  
  
head(df)

## # A tibble: 6 x 12  
## customer\_age Churn churn\_score\_0 CHI\_score Support\_Cases\_0  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 67 0 0 0 0  
## 2 67 0 62 4 0  
## 3 55 0 0 0 0  
## 4 63 0 231 1 1  
## 5 57 0 43 -1 0  
## 6 58 0 138 -10 0  
## # ... with 7 more variables: Support\_Cases\_0\_1 <dbl>, sp\_mon\_0 <dbl>,  
## # SP\_0\_1 <dbl>, logins\_0\_1 <dbl>, blog\_articles\_0\_1 <dbl>,  
## # views\_0\_1 <dbl>, DaysSinceLast\_Login\_0\_1 <dbl>

df <- as.data.frame(df)  
  
cor(df, method = c("pearson", "kendall", "spearman"))

## customer\_age Churn churn\_score\_0  
## customer\_age 1.000000000 0.030215429 0.304893719  
## Churn 0.030215429 1.000000000 -0.084004638  
## churn\_score\_0 0.304893719 -0.084004638 1.000000000  
## CHI\_score -0.151923210 -0.066068588 0.203202136  
## Support\_Cases\_0 -0.075575809 -0.044973034 0.309395401  
## Support\_Cases\_0\_1 -0.010343423 0.005456507 -0.014775639  
## sp\_mon\_0 -0.042218266 -0.054928645 0.376865050  
## SP\_0\_1 -0.008314296 -0.007431734 -0.004129816  
## logins\_0\_1 -0.041608215 -0.042148507 0.317204715  
## blog\_articles\_0\_1 -0.037194849 -0.012889378 0.041695682  
## views\_0\_1 0.003743831 -0.014110047 0.040517359  
## DaysSinceLast\_Login\_0\_1 0.036344153 0.060857901 -0.100888196  
## CHI\_score Support\_Cases\_0 Support\_Cases\_0\_1  
## customer\_age -0.15192321 -0.07557581 -0.010343423  
## Churn -0.06606859 -0.04497303 0.005456507  
## churn\_score\_0 0.20320214 0.30939540 -0.014775639  
## CHI\_score 1.00000000 0.25744080 0.277637350  
## Support\_Cases\_0 0.25744080 1.00000000 0.493100841  
## Support\_Cases\_0\_1 0.27763735 0.49310084 1.000000000  
## sp\_mon\_0 0.22639189 0.65379224 0.254766356  
## SP\_0\_1 0.24109248 0.28800187 0.511297249  
## logins\_0\_1 0.42375781 0.34723551 0.279417586  
## blog\_articles\_0\_1 0.23486945 0.07756877 0.087302531  
## views\_0\_1 -0.01432589 0.02808401 -0.014398444  
## DaysSinceLast\_Login\_0\_1 0.04031129 -0.04270485 0.008042936  
## sp\_mon\_0 SP\_0\_1 logins\_0\_1  
## customer\_age -0.04221827 -0.008314296 -0.04160821  
## Churn -0.05492864 -0.007431734 -0.04214851  
## churn\_score\_0 0.37686505 -0.004129816 0.31720472  
## CHI\_score 0.22639189 0.241092480 0.42375781  
## Support\_Cases\_0 0.65379224 0.288001875 0.34723551  
## Support\_Cases\_0\_1 0.25476636 0.511297249 0.27941759  
## sp\_mon\_0 1.00000000 0.562629623 0.30696452  
## SP\_0\_1 0.56262962 1.000000000 0.21251872  
## logins\_0\_1 0.30696452 0.212518715 1.00000000  
## blog\_articles\_0\_1 0.04776334 0.048583220 0.15185774  
## views\_0\_1 0.02884128 -0.005578740 0.02641103  
## DaysSinceLast\_Login\_0\_1 -0.06274718 0.018643590 -0.02509166  
## blog\_articles\_0\_1 views\_0\_1  
## customer\_age -0.0371948495 0.0037438312  
## Churn -0.0128893776 -0.0141100472  
## churn\_score\_0 0.0416956817 0.0405173590  
## CHI\_score 0.2348694451 -0.0143258906  
## Support\_Cases\_0 0.0775687738 0.0280840124  
## Support\_Cases\_0\_1 0.0873025314 -0.0143984442  
## sp\_mon\_0 0.0477633391 0.0288412849  
## SP\_0\_1 0.0485832201 -0.0055787404  
## logins\_0\_1 0.1518577407 0.0264110337  
## blog\_articles\_0\_1 1.0000000000 0.0002195074  
## views\_0\_1 0.0002195074 1.0000000000  
## DaysSinceLast\_Login\_0\_1 0.0087790889 -0.0029198205  
## DaysSinceLast\_Login\_0\_1  
## customer\_age 0.036344153  
## Churn 0.060857901  
## churn\_score\_0 -0.100888196  
## CHI\_score 0.040311290  
## Support\_Cases\_0 -0.042704851  
## Support\_Cases\_0\_1 0.008042936  
## sp\_mon\_0 -0.062747180  
## SP\_0\_1 0.018643590  
## logins\_0\_1 -0.025091663  
## blog\_articles\_0\_1 0.008779089  
## views\_0\_1 -0.002919820  
## DaysSinceLast\_Login\_0\_1 1.000000000

#install.packages('PerformanceAnalytics')  
library("PerformanceAnalytics")

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':  
##   
## first, last

##   
## Attaching package: 'PerformanceAnalytics'

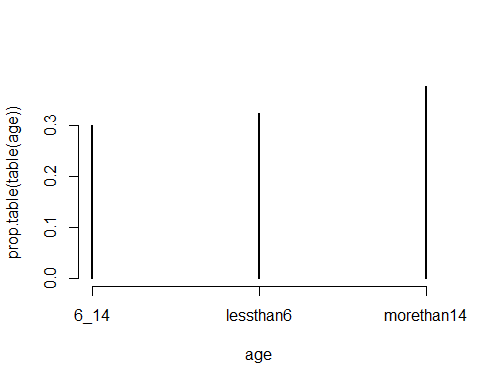
## The following object is masked from 'package:graphics':  
##   
## legend

qwe$ID -> id\_qwe  
df$ID <- id\_qwe

To test that Age is a factor related to Churn rate. We have created 3 levels of age. 1. Less than 6 months old - 32.31% 2. Between 6 and 14 months - 29.96% 3. More than 14 months - 37.71%

Number of customers of age more than 14 months churn more. It shows that there is no dependency on customer age as claimed by Wall’s belief

df$customer\_age -> age  
age <- ifelse((age>=0) & (age<=6),"lessthan6",  
 ifelse((age>6) & (age<=14),"6\_14",  
 ifelse((age>14),"morethan14","NA")))  
  
plot(prop.table(table(age)))



df$age\_level <- age  
str(df)

## 'data.frame': 6347 obs. of 14 variables:  
## $ customer\_age : num 67 67 55 63 57 58 57 46 56 56 ...  
## $ Churn : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ churn\_score\_0 : num 0 62 0 231 43 138 180 116 78 78 ...  
## $ CHI\_score : num 0 4 0 1 -1 -10 -5 -11 -7 -37 ...  
## $ Support\_Cases\_0 : num 0 0 0 1 0 0 1 0 1 0 ...  
## $ Support\_Cases\_0\_1 : num 0 0 0 -1 0 0 1 0 -2 0 ...  
## $ sp\_mon\_0 : num 0 0 0 3 0 0 3 0 3 0 ...  
## $ SP\_0\_1 : num 0 0 0 0 0 0 3 0 0 0 ...  
## $ logins\_0\_1 : num 0 0 0 167 0 43 13 0 -9 -7 ...  
## $ blog\_articles\_0\_1 : num 0 0 0 -8 0 0 -1 0 1 0 ...  
## $ views\_0\_1 : num 0 -16 0 21996 9 ...  
## $ DaysSinceLast\_Login\_0\_1: num 31 31 31 0 31 0 0 6 7 14 ...  
## $ ID : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ age\_level : chr "morethan14" "morethan14" "morethan14" "morethan14" ...

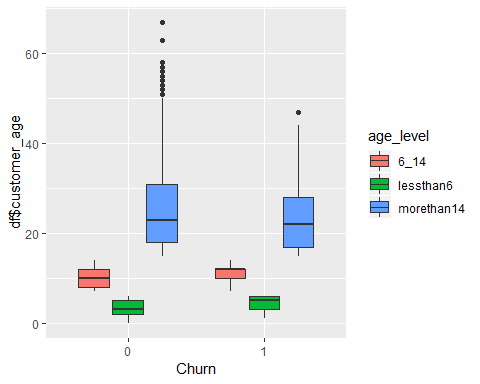
df$age\_level <- as.factor(df$age\_level)  
colnames(df)

## [1] "customer\_age" "Churn"   
## [3] "churn\_score\_0" "CHI\_score"   
## [5] "Support\_Cases\_0" "Support\_Cases\_0\_1"   
## [7] "sp\_mon\_0" "SP\_0\_1"   
## [9] "logins\_0\_1" "blog\_articles\_0\_1"   
## [11] "views\_0\_1" "DaysSinceLast\_Login\_0\_1"  
## [13] "ID" "age\_level"

str(df)

## 'data.frame': 6347 obs. of 14 variables:  
## $ customer\_age : num 67 67 55 63 57 58 57 46 56 56 ...  
## $ Churn : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ churn\_score\_0 : num 0 62 0 231 43 138 180 116 78 78 ...  
## $ CHI\_score : num 0 4 0 1 -1 -10 -5 -11 -7 -37 ...  
## $ Support\_Cases\_0 : num 0 0 0 1 0 0 1 0 1 0 ...  
## $ Support\_Cases\_0\_1 : num 0 0 0 -1 0 0 1 0 -2 0 ...  
## $ sp\_mon\_0 : num 0 0 0 3 0 0 3 0 3 0 ...  
## $ SP\_0\_1 : num 0 0 0 0 0 0 3 0 0 0 ...  
## $ logins\_0\_1 : num 0 0 0 167 0 43 13 0 -9 -7 ...  
## $ blog\_articles\_0\_1 : num 0 0 0 -8 0 0 -1 0 1 0 ...  
## $ views\_0\_1 : num 0 -16 0 21996 9 ...  
## $ DaysSinceLast\_Login\_0\_1: num 31 31 31 0 31 0 0 6 7 14 ...  
## $ ID : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ age\_level : Factor w/ 3 levels "6\_14","lessthan6",..: 3 3 3 3 3 3 3 3 3 3 ...

df$Churn <- as.factor(df$Churn)  
ggplot(aes(y = df$customer\_age, x = Churn, fill = age\_level), data = df) + geom\_boxplot()



###no dependence on age

1. Run a single regression model that best predicts the probability that a customer leaves. What is the predicted probability that customer 672 will leave between December 2011 and February 2012? Is that high or low? Did that customer actually leave? What about customers 354 and 5203?

# Create Training Data  
  
i\_o <- df[which(df$Churn == 1), ] # all 1's  
i\_z <- df[which(df$Churn == 0), ] # all 0's  
set.seed(140) # for repeatability of samples  
  
i\_o\_t\_r <- sample(1:nrow(i\_o), 0.75\*nrow(i\_o)) # 1's for training  
i\_z\_t\_r <- sample(1:nrow(i\_z), 0.75\*nrow(i\_o)) # 0's for training. Pick as many 0's as 1's  
t\_o <- i\_o[i\_o\_t\_r, ]   
t\_z <- i\_z[i\_z\_t\_r, ]  
trainingData <- rbind(t\_o, t\_z) # row bind the 1's and 0's

# Create Test Data  
test\_ones <- i\_o[-i\_o\_t\_r, ]  
test\_zeros <- i\_z[-i\_z\_t\_r, ]  
testData <- rbind(test\_ones, test\_zeros) # row bind the 1's and 0's   
  
dim(testData)

## [1] 5863 14

dim(trainingData)

## [1] 484 14

table(trainingData$Churn)

##   
## 0 1   
## 242 242

table(testData$Churn)

##   
## 0 1   
## 5782 81

set.seed(111)  
Logistic\_regression\_model <- glm(Churn ~ churn\_score\_0 + CHI\_score + sp\_mon\_0+ +Support\_Cases\_0\_1+DaysSinceLast\_Login\_0\_1+views\_0\_1+logins\_0\_1+age\_level+blog\_articles\_0\_1, data=trainingData, family=binomial(link="logit"))  
summary(Logistic\_regression\_model)

##   
## Call:  
## glm(formula = Churn ~ churn\_score\_0 + CHI\_score + sp\_mon\_0 +   
## +Support\_Cases\_0\_1 + DaysSinceLast\_Login\_0\_1 + views\_0\_1 +   
## logins\_0\_1 + age\_level + blog\_articles\_0\_1, family = binomial(link = "logit"),   
## data = trainingData)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.83111 -1.05898 0.03867 1.08486 2.18980   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 9.178e-01 2.199e-01 4.173 3.01e-05 \*\*\*  
## churn\_score\_0 -7.460e-03 1.967e-03 -3.792 0.000149 \*\*\*  
## CHI\_score -4.846e-03 4.057e-03 -1.195 0.232255   
## sp\_mon\_0 -1.615e-02 9.244e-02 -0.175 0.861301   
## Support\_Cases\_0\_1 1.408e-01 8.490e-02 1.659 0.097208 .   
## DaysSinceLast\_Login\_0\_1 1.482e-02 5.000e-03 2.964 0.003034 \*\*   
## views\_0\_1 -9.647e-05 9.016e-05 -1.070 0.284628   
## logins\_0\_1 1.878e-03 2.715e-03 0.692 0.489186   
## age\_levellessthan6 -1.400e+00 2.849e-01 -4.914 8.94e-07 \*\*\*  
## age\_levelmorethan14 -1.652e-01 2.238e-01 -0.738 0.460561   
## blog\_articles\_0\_1 8.410e-03 6.576e-02 0.128 0.898238   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 670.97 on 483 degrees of freedom  
## Residual deviance: 606.18 on 473 degrees of freedom  
## AIC: 628.18  
##   
## Number of Fisher Scoring iterations: 4

anova(Logistic\_regression\_model)

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Churn  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 483 670.97  
## churn\_score\_0 1 14.6063 482 656.36  
## CHI\_score 1 2.6804 481 653.68  
## sp\_mon\_0 1 0.4702 480 653.21  
## Support\_Cases\_0\_1 1 6.3982 479 646.81  
## DaysSinceLast\_Login\_0\_1 1 11.2583 478 635.55  
## views\_0\_1 1 2.1069 477 633.45  
## logins\_0\_1 1 0.0000 476 633.45  
## age\_level 2 27.2543 474 606.19  
## blog\_articles\_0\_1 1 0.0164 473 606.18

colnames(trainingData)

## [1] "customer\_age" "Churn"   
## [3] "churn\_score\_0" "CHI\_score"   
## [5] "Support\_Cases\_0" "Support\_Cases\_0\_1"   
## [7] "sp\_mon\_0" "SP\_0\_1"   
## [9] "logins\_0\_1" "blog\_articles\_0\_1"   
## [11] "views\_0\_1" "DaysSinceLast\_Login\_0\_1"  
## [13] "ID" "age\_level"

colnames(qwe)

## [1] "ID" "customer\_age"   
## [3] "Churn" "churn\_score\_0"   
## [5] "CHI\_score" "Support\_Cases\_0"   
## [7] "Support\_Cases\_0\_1" "sp\_mon\_0"   
## [9] "SP\_0\_1" "logins\_0\_1"   
## [11] "blog\_articles\_0\_1" "views\_0\_1"   
## [13] "DaysSinceLast\_Login\_0\_1"

predicted <- plogis(predict(Logistic\_regression\_model, testData)) # predicted scores

The Actual probabilities of the required customers are given below. It is low and less than 0.5. No, these customers did not left actually. customer 354 - 0.4809895 customer 672 - 0.46486432 customer 5203 - 0.32574207

library(InformationValue)

##   
## Attaching package: 'InformationValue'

## The following objects are masked from 'package:caret':  
##   
## confusionMatrix, precision, sensitivity, specificity

optCutOff <- 0.75

require(MASS)

## Loading required package: MASS

require(car)

## Loading required package: car

## Loading required package: carData

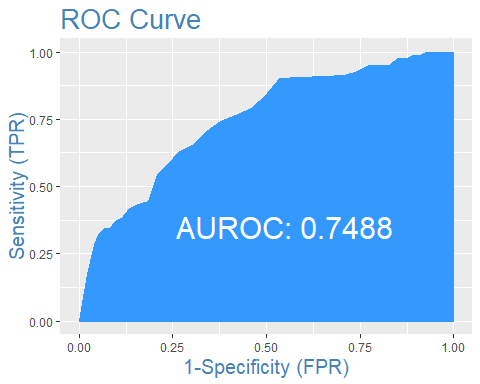
vif(Logistic\_regression\_model)

## GVIF Df GVIF^(1/(2\*Df))  
## churn\_score\_0 1.523989 1 1.234499  
## CHI\_score 1.455768 1 1.206552  
## sp\_mon\_0 1.436344 1 1.198476  
## Support\_Cases\_0\_1 1.386437 1 1.177470  
## DaysSinceLast\_Login\_0\_1 1.033686 1 1.016704  
## views\_0\_1 1.019078 1 1.009494  
## logins\_0\_1 1.475719 1 1.214792  
## age\_level 1.330097 2 1.073917  
## blog\_articles\_0\_1 1.271671 1 1.127684

##devtools::install\_github("InformationValue")  
misClassError(testData$Churn, predicted, threshold = optCutOff)

## [1] 0.0539

plotROC(testData$Churn,predicted)



table(testData$Churn)

##   
## 0 1   
## 5782 81

Concordance(testData$Churn, predicted)

## $Concordance  
## [1] 0.7480388  
##   
## $Discordance  
## [1] 0.2519612  
##   
## $Tied  
## [1] 5.551115e-17  
##   
## $Pairs  
## [1] 468342

sensitivity(testData$Churn, predicted, threshold = optCutOff)

## [1] 0.3209877

#> 0.3089  
specificity(testData$Churn, predicted, threshold = optCutOff)

## [1] 0.9548599

confusionMatrix(testData$Churn, predicted, threshold = optCutOff)

## 0 1  
## 0 5521 55  
## 1 261 26

df[c(354,672,5203),]

## customer\_age Churn churn\_score\_0 CHI\_score Support\_Cases\_0  
## 354 13 0 139 -29 0  
## 672 16 0 148 1 0  
## 5203 4 0 37 32 1  
## Support\_Cases\_0\_1 sp\_mon\_0 SP\_0\_1 logins\_0\_1 blog\_articles\_0\_1  
## 354 0 0 0 -4 1  
## 672 0 0 0 17 1  
## 5203 1 0 0 0 0  
## views\_0\_1 DaysSinceLast\_Login\_0\_1 ID age\_level  
## 354 244 -1 354 6\_14  
## 672 85 2 672 morethan14  
## 5203 1 5 5203 lessthan6

class(predicted)

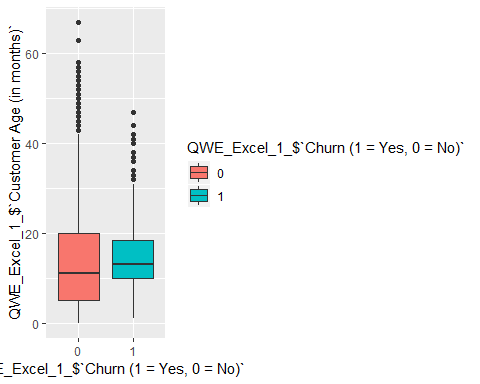
## [1] "numeric"

predicted <- as.data.frame(predicted)  
colnames(predicted)

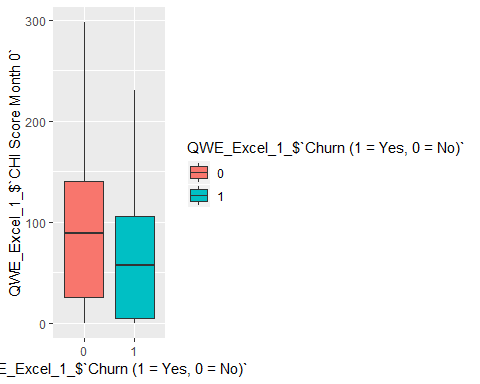
## [1] "predicted"

trainingData$ID -> id\_train  
testData$ID -> id\_test  
predicted$id <- id\_test  
View(predicted)

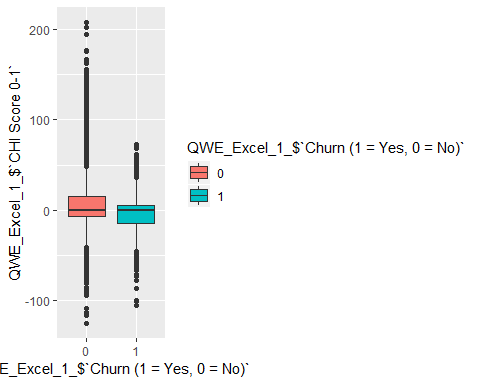
##factors responsible for customers leaving out  
library(magrittr)  
QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)` <- as.factor(QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)  
QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`Customer Age (in months)`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



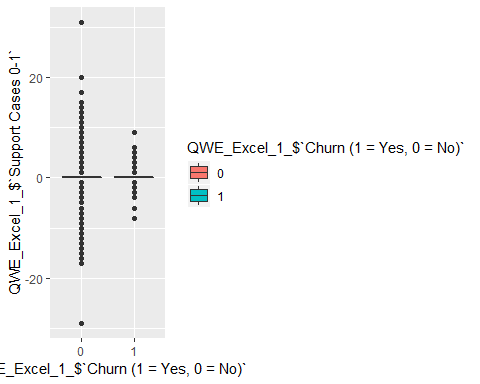
QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`CHI Score Month 0`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



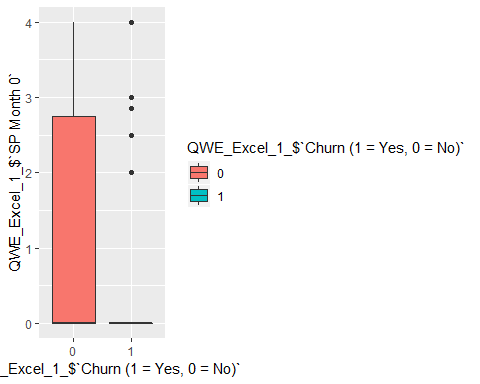
QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`CHI Score 0-1`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



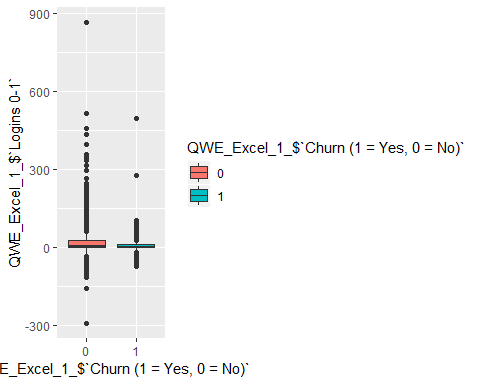
QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`Support Cases 0-1`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



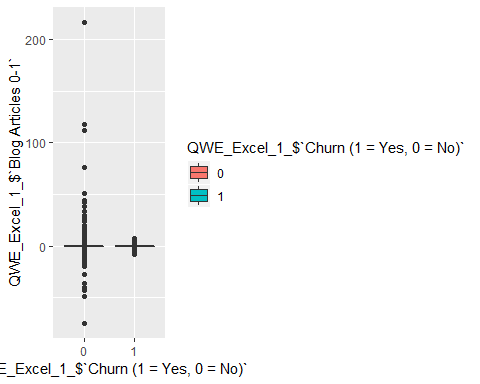
QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`SP Month 0`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



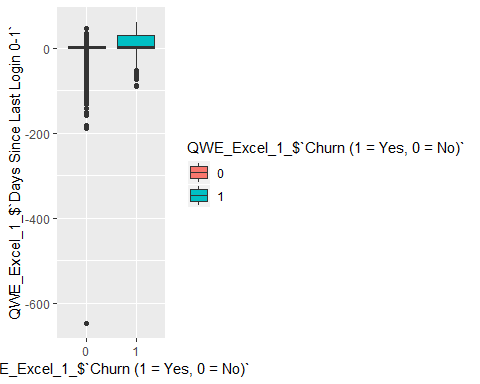
QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`Logins 0-1`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`Blog Articles 0-1`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()



QWE\_Excel\_1\_ %>% ggplot(aes(x = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`, y = QWE\_Excel\_1\_$`Days Since Last Login 0-1`, fill = QWE\_Excel\_1\_$`Churn (1 = Yes, 0 = No)`)) +geom\_boxplot()

 Plot 1 - Longevity in months has very less impact on the churn rate of customers. Plot 2 - CHI score for month 0 shows that customer who have churned have lower CHI score relative to other customers. Plot 3 - CHI score for month 0-1 shows that customer who have churned have lower CHI score relative to other customers. Plot 4 - Plot 5 - Those customers who have high average Support Priority. will not churn more frequently. Plot 6 - Those customers who have churned have lesser number of logins as compared to other customers. Plot 7 - Plot 8 - Days since last login is also a factor responsible for Churn rate of customers. From Boxplot , we can see that the customers who have churned have more days since last login.

1. How sensible is the approach with a single model? Can you suggest a better approach? Provide updated estimates of probabilities that customers 672, 354 and 5,203 will leave.

Lets try out with SVM model

###SVM model###  
library("e1071")

##   
## Attaching package: 'e1071'

## The following objects are masked from 'package:PerformanceAnalytics':  
##   
## kurtosis, skewness

colnames(testData)

## [1] "customer\_age" "Churn"   
## [3] "churn\_score\_0" "CHI\_score"   
## [5] "Support\_Cases\_0" "Support\_Cases\_0\_1"   
## [7] "sp\_mon\_0" "SP\_0\_1"   
## [9] "logins\_0\_1" "blog\_articles\_0\_1"   
## [11] "views\_0\_1" "DaysSinceLast\_Login\_0\_1"  
## [13] "ID" "age\_level"

colnames(trainingData)

## [1] "customer\_age" "Churn"   
## [3] "churn\_score\_0" "CHI\_score"   
## [5] "Support\_Cases\_0" "Support\_Cases\_0\_1"   
## [7] "sp\_mon\_0" "SP\_0\_1"   
## [9] "logins\_0\_1" "blog\_articles\_0\_1"   
## [11] "views\_0\_1" "DaysSinceLast\_Login\_0\_1"  
## [13] "ID" "age\_level"

trainingData$ID -> id\_train  
testData$ID -> id\_test  
trainingData <- trainingData[,c(1,2,3,4,5,6,7,8,9,10,11,12)]  
testData <- testData[,c(1,2,3,4,5,6,7,8,9,10,11,12)]  
svm\_model <- svm(as.numeric(as.character(Churn)) ~ ., data=trainingData)  
  
class(svm\_model)

## [1] "svm.formula" "svm"

pred <- plogis(predict(svm\_model,testData))

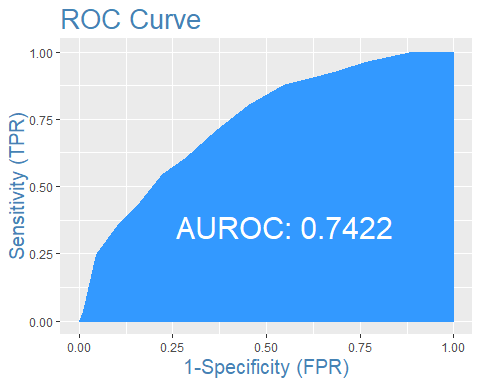
#354 - 0.744231670  
#672 - 0.622208856  
#5203 - 0.18854133  
optCutOff <- .75  
  
misClassError(testData$Churn, pred, threshold = optCutOff)

## [1] 0.0155

str(pred)

## Named num [1:5863] 0.669 0.619 0.609 0.527 0.706 ...  
## - attr(\*, "names")= chr [1:5863] "94" "105" "156" "204" ...

plotROC(as.numeric(as.character(testData$Churn)),as.numeric(pred))



Concordance(testData$Churn, pred)

## $Concordance  
## [1] 0.7404034  
##   
## $Discordance  
## [1] 0.2595966  
##   
## $Tied  
## [1] 0  
##   
## $Pairs  
## [1] 468342

sensitivity(testData$Churn, pred, threshold = optCutOff)

## [1] 0

# 0.50  
specificity(testData$Churn, pred, threshold = optCutOff)

## [1] 0.9982705

#0.822  
confusionMatrix(testData$Churn, pred, threshold = optCutOff)

## 0 1  
## 0 5772 81  
## 1 10 0

View(pred)

Below are the given probabilities for the customers. customer 354 - 0.6098852 customer 672 - 0.5867321 customer 5203 - 0.5405019

This time, we have implemented SVM model which means that a kernel trick is used to transformation and then based on these transformations, it finds an optimal boundary between the possible outputs. This technique is used for the customers to find an optimal hyperplane to classify the churned customers and non- churn customers.

These customers have higher probability to churn using SVM model but as per the cutoff , it is still lesser than 0.75