IDS 572 SPRING 2020

HOMEWORK 2

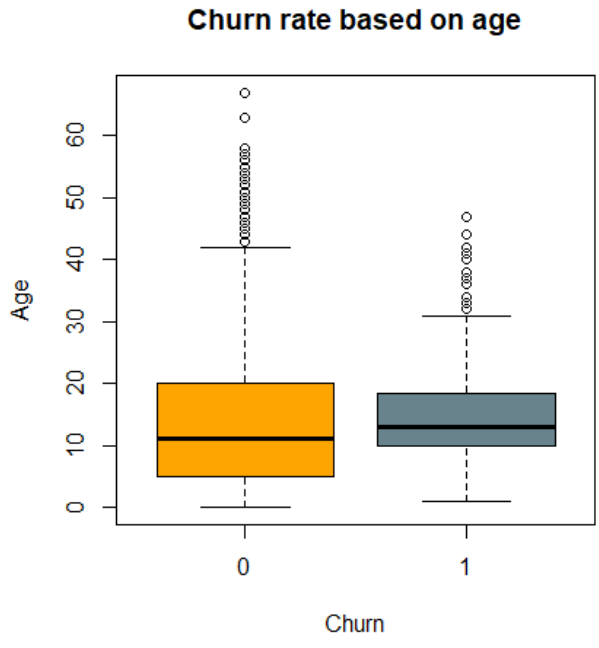
- RACHANA MANJUNATH (rmanju2)

- SOWMYA SANKRANTHI (ssankr2)

- SHREYA GOWDA (sgowda6)

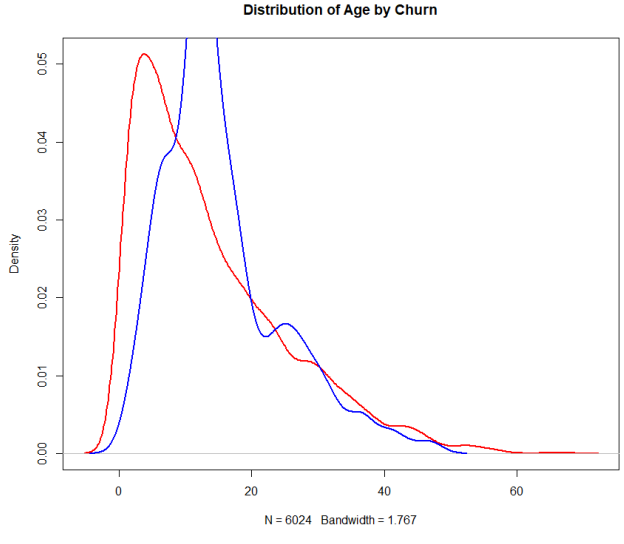
- ***Solution 1***

```
library(ISLR)
install.packages("dplyr")
library(car)
library(corrplot) # plot correlations
library(dplyr) # data aggregates
library(Hmisc) # for correlation test of multiple variables
library(gplots)
library(ggplot2)
library("readxl")
cust_stats <- read_excel(file.choose(), sheet = 2)
str(cust_stats)
colnames(cust_stats) <- c("ID" , "Age" , "Churn" , "CHIScore_Month0" , "CHIScore_0_to_1" , "SupportCases_Month0" , "SupportCases_0_to_1" , "SP_Month0" , "SP_0_to_1" , "Logins_0_to_1" , "BlogArticles_0_to_1" , "Views_0_to_1" , "Days_Since_LastLogin_0_to_1")  #rename
cust_stats$Churn <- as.factor(cust_stats$Churn)
 boxplot( cust_stats$Age ~ cust_stats$Churn, data=cust_stats, main="Churn rate based on age",
        xlab="Churn", ylab="Age", col=c("orange", "lightblue4"))
churn_0 <- subset(cust_stats, cust_stats$Churn==0) #6024
churn_1 <- subset(cust_stats, cust_stats$Churn==1) #323
 fivenum(sort(churn_0$Age)) # 0  5 11 20 67
quantile(sort(churn_0$Age))
 fivenum(sort(churn_1$Age)) # 1.0 10.0 13.0 18.5 47.0
```



We can see from the boxplot there are a few customers with age 5-11 months who do not churn and few customers with age 10-13 months who do churn. There is some overlap between customer who churn and customers who don't. Hence, we could say that there is no significant/apparent relationship between age and churn (0-NO CHURN and 1-CHURN) when age is considered independently. But however, later on, we do see that age along with other variables does have an impact on prediction of churn.

In addition, customers with age width 10-20 months are the riskiest customers who might tend to leave when compared to newer and older age groups. The same thing can be observed below:



- *Solution 2*

Logistic Regression Model:

```
set.seed(300)

data <- cust_stats[-c(354,672,5203),]

indx <- sample(2, nrow(data), replace = TRUE, prob = c(0.7,0.3))

train_data <- data[indx == 1,]

test_data <- data[indx == 2 ,]

test_data <- rbind(test_data,cust_stats[c(354,672,5203),])

logitModel <- glm(Churn ~ ., data = train_data, family = "binomial") #a logistic regression model for target variable churn, including all the other variables in the data set

summary(logitModel) #we see that most variables have a very high p value and hence are not significant. So we will exclude these variables and include just Days_Since_LastLogin_0-1,CHIScore_0-1,CHIScore_Month0,Age, CHIScore_Month0, Views_0_to_1.

logitModel_new <- glm(Churn ~ Age + CHIScore_Month0 + CHIScore_0_to_1 + Views_0_to_1 + Days_Since_LastLogin_0_to_1, data = train_data, family = "binomial")

summary(logitModel_new)

options(scipen = 99)

Pred <- predict(logitModel_new,type="response")

Pred

options(max.print=1000000)
```

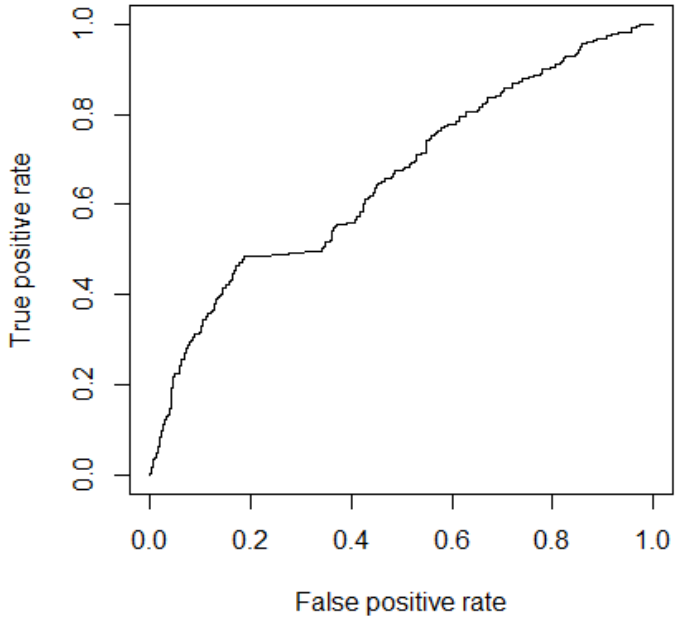
Plot the ROC Curve to find the threshold/cut off point

```
library(ROCR)
ROC_pred <- prediction(Pred,train_data$Churn)

perf <- performance(ROC_pred, "tpr", "fpr")

plot(perf)

auc <- performance(ROC_pred, "auc")

auc

auc <- unlist(slot(auc, "y.values"))

auc

# How to find the best cut-off point in ROC curve
# The performance() function for ROC curve returns
```

```
# tpr, fpr and alpha-values (cut-off points). We need
# to write a function that receives these information and
# returns the best cut-off point
# Hence the input argument to the following function is perf

opt.cut <- function(perf)

 cut.ind <- mapply(FUN = function(x,y,p){d=(x-0)^2+(y-1)^2 # We compute the distance of all the points
from the corner point [1,0]

  ind<- which(d==min(d)) # We find the index of the point that is closest to the corner

  c(recall = y[[ind]], specificity = 1-x[[ind]],cutoff = p[[ind]])},perf@x.values,
perf@y.values,perf@alpha.values)


 print(opt.cut(perf)) #0.06522078 ~ 0.064 is the OPTIMUM threshold
```

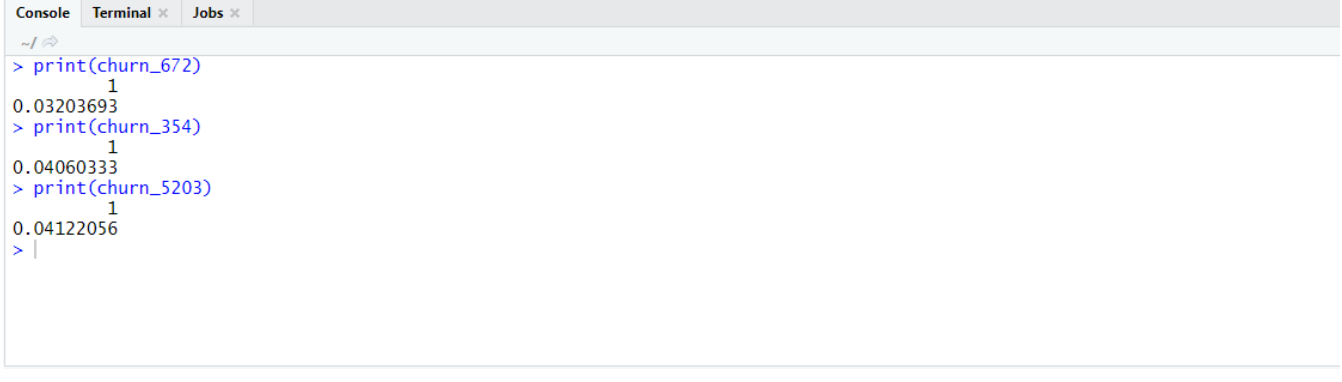
0.064 is the OPTIMUM threshold

```
Console  Terminal  Jobs
~/
> print(opt.cut(perf))
                [,1]
recall      0.48416290
specificity 0.81257616
cutoff      0.06522078
>
```



To predict churn probabilities of Customers with ID 672,354,5203

```
logistic_Pred <- predict(logitModel_new, newdata = test_data, type = "response")

logistic_Pred

View(logistic_Pred)

#354 672 5203

Class <- ifelse(Pred > 0.064, "YES", "NO")

Class

data1 <- data.frame(cust_stats[cust_stats$ID==672,])
churn_672 <- logitModel_new %>% predict(data1, type = "response")

data2 <- data.frame(cust_stats[cust_stats$ID==354,])
churn_354 <- logitModel_new %>% predict(data2, type = "response")

data3 <- data.frame(cust_stats[cust_stats$ID==5203,])
churn_5203 <- logitModel_new %>% predict(data3, type = "response")


print(churn_672)
print(churn_354)
print(churn_5203)
```



a) **The predicted probability that Customer 672 will leave between December 2011 and February 2012 is 0.03203693, which is lesser than our threshold 0.064. Therefore, churn is NO, the customer with ID 672 will not churn. From the actual dataset, the churn value for ID 672 is 0, the customer does not churn. Hence, our prediction is right.**
b) **The predicted probability that Customer 354 will leave between December 2011 and February 2012 is 0.04060333, which is lesser than our threshold 0.064. Therefore, churn is NO, the customer with ID 672 will not churn. From the actual dataset, the churn value for ID 354 is 0, the customer does not churn. Hence, our prediction is right.**

c) The predicted probability that Customer 5203 will leave between December 2011 and February 2012 is 0.04122056, which is lesser than our threshold 0.064. Therefore, churn is NO, the customer with ID 5203 will not churn. From the actual dataset, the churn value for ID 354 is 0, the customer does not churn. Hence, our prediction is right.

Accuracy of the model

```
cf <- table(test_data$Churn,logistic_Pred>0.064)

cf

(cf[1] + cf[4])/sum(cf)
```

We find that our model has an accuracy of 0.7577855 i.e. 75.77%. The confusion matrix is as below:

	FALSE	**TRUE**
0	**1487**	**434**
1	**56**	**46**

Solution 3:

List of 100 customers with the highest churn probabilities and the top three drivers of churn for each customer.

```
#top 100 customers with highest probability of churn

logistic_Pred_top100 <- predict(logitModel_new, newdata = cust_stats, type = "response")

logistic_Pred_top100

d<-sort(logistic_Pred_top100, decreasing = TRUE) #to obtain 100 customers with highest probabilities of churning

d

max_100 <- head(d,100)
library(xlsx)
write.csv(max_100, "C:\\Users\\rachn\\Desktop\\Data Mining\\mydata.csv")
```

It is as below as well:

CustomerID	Probability of Churn	Churn (YES/NO)
2287	0.675299096	YES
357	0.624962467	YES
929	0.374525567	YES
109	0.32407304	YES
1971	0.262189743	YES
2025	0.259431618	YES
1	0.2212989	YES
2076	0.214519693	YES
76	0.212521123	YES
1363	0.20235525	YES
586	0.199406079	YES
14	0.196995676	YES
18	0.192434087	YES
3	0.192384301	YES
299	0.192311467	YES
2244	0.191259901	YES
1287	0.19084839	YES
21	0.190109641	YES
1929	0.188032314	YES
884	0.186765248	YES
1459	0.182890951	YES
1520	0.182791743	YES
51	0.178127957	YES
2546	0.177272667	YES
2913	0.176633665	YES
183	0.175694923	YES
1672	0.17515398	YES
2680	0.173454884	YES
128	0.173057438	YES
59	0.172651603	YES
1286	0.172412344	YES
55	0.1719533	YES
2240	0.171094872	YES
1236	0.170901916	YES
2599	0.170685103	YES
121	0.17056098	YES
2922	0.170110682	YES
1021	0.170047705	YES
1862	0.169677783	YES
137	0.168490516	YES
335	0.168243856	YES

2080	0.168081931	YES
1143	0.167599075	YES
2481	0.167428203	YES
3604	0.167306164	YES
123	0.167225942	YES
3340	0.166733831	YES
1574	0.166626732	YES
154	0.16650683	YES
68	0.166440143	YES
2951	0.166419199	YES
146	0.165075147	YES
1616	0.164499535	YES
119	0.164409787	YES
171	0.164409787	YES
190	0.16438778	YES
89	0.163455899	YES
1141	0.163127424	YES
2838	0.162720473	YES
2289	0.16233313	YES
101	0.162290447	YES
42	0.161147181	YES
5	0.159057032	YES
95	0.159007826	YES
61	0.158775293	YES
2	0.156573024	YES
2924	0.156501131	YES
1438	0.156486992	YES
1392	0.156375218	YES
139	0.156232622	YES
3671	0.15583272	YES
156	0.155710265	YES
2335	0.15468057	YES
106	0.154444892	YES
4245	0.153966018	YES
1393	0.152921826	YES
203	0.151786814	YES
2255	0.150853695	YES
57	0.150754869	YES
1395	0.150751119	YES
1478	0.150751119	YES
2235	0.150751119	YES
69	0.150406704	YES
798	0.150346604	YES

3124	0.150107553	YES
1204	0.14965454	YES
1151	0.149620199	YES
142	0.149227859	YES
1488	0.148644096	YES
1693	0.148488821	YES
2830	0.148186923	YES
2739	0.147738266	YES
4191	0.14772045	YES
3258	0.147152872	YES
2296	0.14702406	YES
5314	0.146790506	YES
2242	0.145923277	YES
62	0.14591748	YES
3042	0.145127967	YES
2903	0.144719787	YES

We have exported our list of 100 customers with highest probability of churn to an excel file. Irrespective of the fact that IF these customers actually churned in the last two months, our model predicts their potential of churn in the following months. QWE Inc. could be more cautious about these customers and reach out to them in order to retain them in the coming months.

Also, an important point to note is what drives these customers to quit. According to our logistic regression model, the variables of highest significance (with lowest p value) are: CHIScore_Month0, CHIScore_0_to_1 and Views_0_to_1.