Medical Insurance Prediction using Logistic Regression

KARIUKI REUBEN MWANGI

5/29/2022

## Introduction

In this analysis,I am developing a predictive model that will be used to predict the number of people who are likely to purchase medical insurance in a given day or reject the medical insurance. Hence, those who purchased, and not purchased the medical insurance.

This is a model which is used mostly in predictive analytics scenario.

### Importing data

# loading dataset in R for analysis  
medical\_insur <- read.csv("C:/Users/kariuki Reuben/Downloads/Medical\_insurance\_dataset - Medical\_insurance\_dataset.csv")  
  
# loading necessary library for this project  
library("tidyverse")

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

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

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

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

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

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

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

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("data.table")

## Warning: package 'data.table' was built under R version 4.0.5

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

library("dplyr")  
library("gridExtra")

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

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library("scales")

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

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

# class of the dataset  
  
class(medical\_insur)# identifying the class of the dataset

## [1] "data.frame"

names(medical\_insur)# exploring all the available attributes in the dataset

## [1] "User.ID" "Gender" "Age" "EstimatedSalary"  
## [5] "Purchased"

## Descriptive Analytics of the dataset

There is need to carry out the encoding in the variables to make it more easy when coming up with predictive model.

glimpse(medical\_insur) # the structure of the dataset

## Rows: 400  
## Columns: 5  
## $ User.ID <int> 15624510, 15810944, 15668575, 15603246, 15804002, 1572~  
## $ Gender <chr> "Male", "Male", "Female", "Female", "Male", "Male", "F~  
## $ Age <int> 19, 35, 26, 27, 19, 27, 27, 32, 25, 35, 26, 26, 20, 32~  
## $ EstimatedSalary <int> 19000, 20000, 43000, 57000, 76000, 58000, 84000, 15000~  
## $ Purchased <chr> "not-purchased", "not-purchased", "not-purchased", "no~

summary(medical\_insur)# summary statistics of the dataset

## User.ID Gender Age EstimatedSalary   
## Min. :15566689 Length:400 Min. :18.00 Min. : 15000   
## 1st Qu.:15626764 Class :character 1st Qu.:29.75 1st Qu.: 43000   
## Median :15694342 Mode :character Median :37.00 Median : 70000   
## Mean :15691540 Mean :37.66 Mean : 69743   
## 3rd Qu.:15750363 3rd Qu.:46.00 3rd Qu.: 88000   
## Max. :15815236 Max. :60.00 Max. :150000   
## Purchased   
## Length:400   
## Class :character   
## Mode :character   
##   
##   
##

From the summary statistics there is clear evidence that data is fit to carry out model training since no presence of outliers in the dataset.

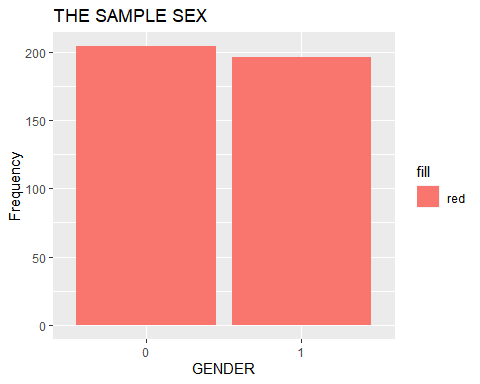
## Future Encoding

medical\_insur$Purchased= (factor(medical\_insur$Purchased,  
 levels = c('not-purchased', 'purchased'), labels = c(0, 1)))  
  
medical\_insur$Gender = (factor(medical\_insur$Gender,levels = c("Female","Male"),labels = c(0,1)))  
  
  
sum(is.na(medical\_insur))# no presence of missing values and the data is ready for visualization and model building.

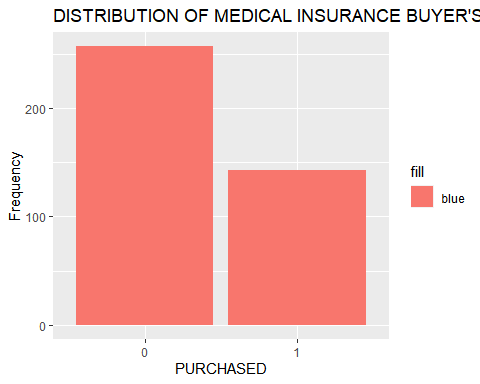
## [1] 0

### Data Visualization 1

# data visualization  
  
sex<- medical\_insur %>% group\_by(Gender )   
ggplot(sex) + geom\_bar(aes(x=Gender,fill = "red")) + labs(title="THE SAMPLE SEX") + xlab("GENDER") + ylab("Frequency")



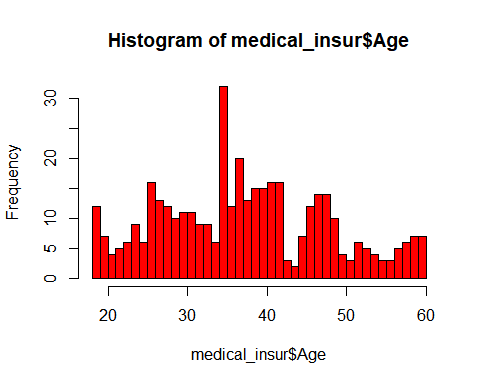
purchased<- medical\_insur%>% group\_by(Purchased)   
ggplot(purchased) + geom\_bar(aes(x=Purchased ,fill = "blue")) + labs(title="DISTRIBUTION OF MEDICAL INSURANCE BUYER'S") + xlab("PURCHASED") + ylab("Frequency")

 -**Answer** + From the visualization the total number of Females are more than Male in the dataset. This will assist us in identifying which gender is more likely to purchase the medical insurance.

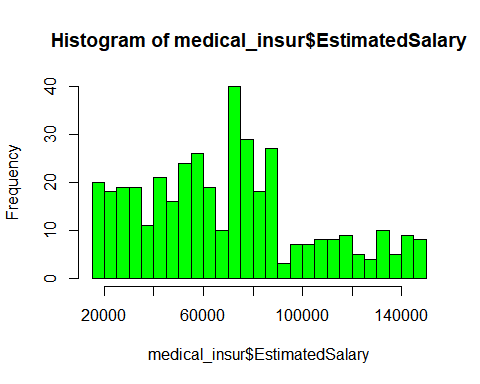
* From the second chart, the total number of people not purchasing the medical insurance are more than those purchasing medical insurance.
* There is need of more aggressive marketing to ensure that the company gets more people purchasing the medical insurance.

### Data Visualization 2

set.seed(123)# to make my work reproducedable  
  
hist(medical\_insur$Age,breaks = 30,col = "red")# there is a normal distribution in the age column



hist(medical\_insur$EstimatedSalary,breaks = 30 , col = "green")# the dataset is normally distributed.



-**Answer** + From the first chart, the distribution of Age column is normal distribution, with the dorminant age in the range of 30-40 years.

* From the second chart, the distribution of Salary column is normal with the maximum salary between 50,000 and 90,0000.

### Model Building

# model building  
library(caTools)

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

medical\_pred = sample.split(medical\_insur$Purchased,SplitRatio = 0.80)  
  
train\_data = subset(medical\_insur,medical\_pred ==TRUE)  
  
test\_data = subset(medical\_insur,medical\_pred == FALSE)  
  
  
# Run GLM Model with one variable  
  
logistic\_1<- glm( Purchased~., family = "binomial", data = train\_data)  
summary(logistic\_1)# summary statistics of the model

##   
## Call:  
## glm(formula = Purchased ~ ., family = "binomial", data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1120 -0.4859 -0.1094 0.3014 2.5172   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.555e+01 3.836e+01 -0.406 0.685   
## User.ID 7.975e-08 2.445e-06 0.033 0.974   
## Gender1 3.464e-01 3.555e-01 0.974 0.330   
## Age 2.660e-01 3.386e-02 7.855 4.00e-15 \*\*\*  
## EstimatedSalary 4.158e-05 6.691e-06 6.214 5.16e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 416.79 on 319 degrees of freedom  
## Residual deviance: 208.87 on 315 degrees of freedom  
## AIC: 218.87  
##   
## Number of Fisher Scoring iterations: 6

-**Answer**

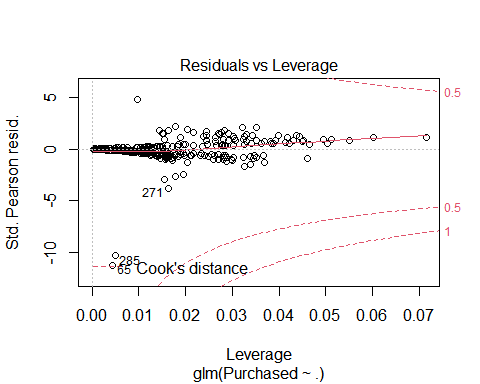
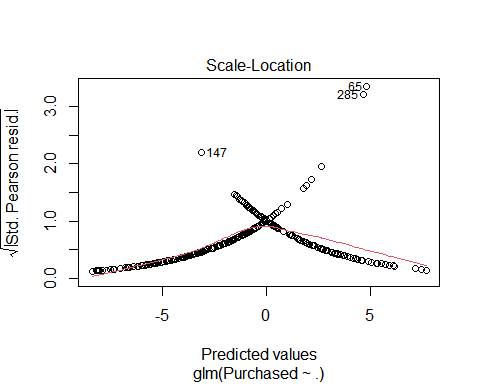
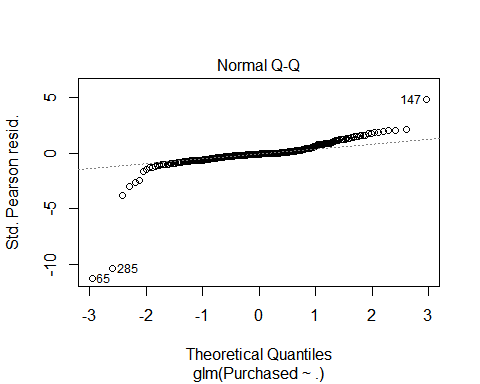
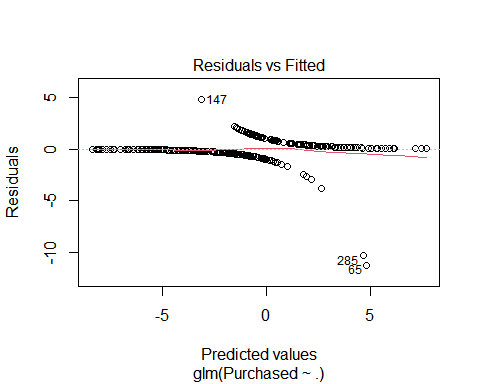
* Gender and ID number are not signficant variable in predicting the number of people to purchase the medical insurance
* One of the most significant variable to consinder is the Age and Estimated Salary
* This might open up our mind in determing the correct age bracket to target while marketing and selling these medical insurance.
* Also we have a hint of the right salary bracket of these people purchasing the medical insurance and those not buying the medical cover.

### PLOT the logistic model

-**Answer**

* Provide the assumption of logistic model

# plot the logistic model outputs  
  
plot(logistic\_1)



### Developing prediction from the model

-**Answer**

* testing how the logistic model will perform on a new dataset

# testing the model using a new data   
predict\_medical <- predict(logistic\_1, newdata=test\_data, type = 'response')  
  
# Construct a confusion matrix for the prdicted values  
  
y\_predict<-table(test\_data$Purchased, predict\_medical > 0.5)  
y\_predict

##   
## FALSE TRUE  
## 0 44 7  
## 1 8 21

# accuaracy of the model will be  
  
Accuracy = (44+21)/(44+7+8+21)  
  
Accuracy

## [1] 0.8125

# model is perfoming well and given us a prediction accuaracy of 81.25%  
  
# the baseline accuracy is 0.8125= 81.25%  
dimnames(y\_predict)[[1]]=c("A","B")  
dimnames(y\_predict)[[2]]= c("A","B")  
y\_predict

##   
## A B  
## A 44 7  
## B 8 21

##code for specificity $ sensitivity ,and accuracy  
library(caret)

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

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

confusionMatrix(y\_predict)

## Confusion Matrix and Statistics  
##   
##   
## A B  
## A 44 7  
## B 8 21  
##   
## Accuracy : 0.8125   
## 95% CI : (0.7097, 0.8911)  
## No Information Rate : 0.65   
## P-Value [Acc > NIR] : 0.001109   
##   
## Kappa : 0.5913   
##   
## Mcnemar's Test P-Value : 1.000000   
##   
## Sensitivity : 0.8462   
## Specificity : 0.7500   
## Pos Pred Value : 0.8627   
## Neg Pred Value : 0.7241   
## Prevalence : 0.6500   
## Detection Rate : 0.5500   
## Detection Prevalence : 0.6375   
## Balanced Accuracy : 0.7981   
##   
## 'Positive' Class : A   
##

### Plotting the ROCR CURVE

#ROCR Curve  
#install.packages("ROCR")  
library(ROCR)

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

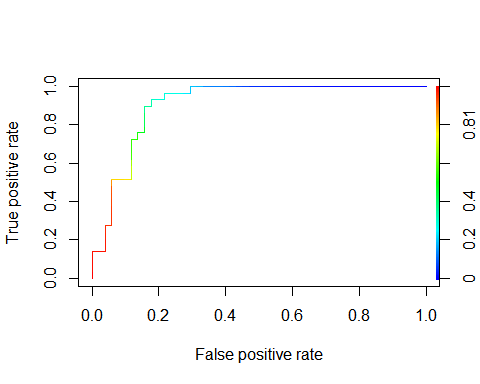
# identifying the prediction instance  
ROCRpred <- prediction(predict\_medical, test\_data$Purchased)  
  
ROCRpred

## A prediction instance  
## with 80 data points

# identifying the perfomance instance  
perf <- performance(ROCRpred,"tpr","fpr")  
perf

## A performance instance  
## 'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')  
## with 81 data points

# plotting the ROCR CURVE  
  
plot(perf,colorize = TRUE)

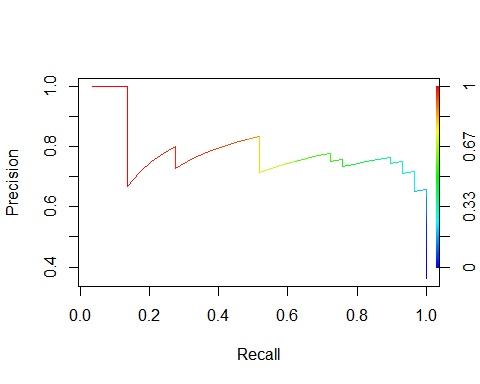


### Precision Curve

# performance instance  
perf <- performance(ROCRpred, "prec", "rec")  
perf

## A performance instance  
## 'Recall' vs. 'Precision' (alpha: 'Cutoff')  
## with 81 data points

# the plot of precision  
  
plot(perf,colorize = TRUE)

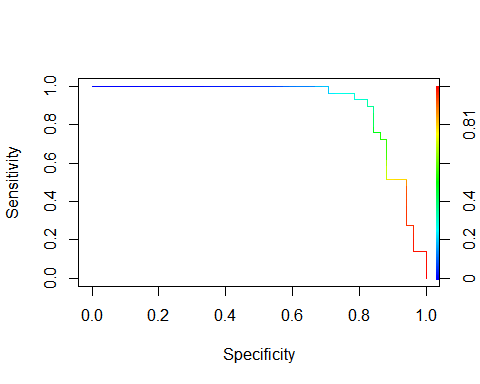


### Plottin the Sensitivity Curve & Specificity

# sensitivity/specificity curve (x-axis: specificity,  
# y-axis: sensitivity)  
perf <- performance(ROCRpred, "sens", "spec")  
perf

## A performance instance  
## 'Specificity' vs. 'Sensitivity' (alpha: 'Cutoff')  
## with 81 data points

# plotting the curve  
  
plot(perf,colorize = TRUE)



## Welcome to the End of my Analysis

#### Munene