Unsupervised Learning:Cryptography IP, R Basics ,

Elvis Njoroge

2022-03-21

{r setup, include=FALSE} knitr::opts\_chunk$set(echo = TRUE)

## R Markdown

####Intoduction An analysis of the variables that affect the ability of an Individual to click ads, Elvis Njoroge , Rbasics.

####Defining the Question’ Technology has revolutionized numerous fields including advertising. Digital marketing has explored numerous pockets, taking advantage of media tools such as blogs and online news bulletins. Cryptography puts a publisher in control of their media content allowing them to explore its economic benefits. However, content published may have different target audience prompting a rigorous research. A blogger, interested in expanding her revenue streams employed the services of a Data Scientist t investigate the preference of her audience when clicking ads.

#####Defining the context The entrepreneur wants to expand her target audience by narrowing down content that encourages her audience to click on ads.

####Metrcs of Succeess An analysis conducted will provide some insight on the relationship between variables offering recommendations.

####Data Relevance

The data being used has been collected and compiled by a trusted source.(<http://bit.ly/IPAdvertisingData>)

####Loading the Dataset

#importing the dataset

#importing the dataset  
  
vert<- read.csv("http://bit.ly/IPAdvertisingData")  
  
#preview of the advert dataset, the first 6 rows  
  
head(vert)

## Including Plots

You can also embed plots, for example:

##Importing the libraries needed  
library(tidyverse)  
library(data.table)  
library(ggplot2)  
library(lattice)  
library(caret)  
library(rpart)  
library(dplyr)  
library(epiDisplay)  
library(rpart.plot)  
library(fastAdaboost)

#Checking the number of rows and columns  
dim(vert)

The dataset contains 1000 rows and 10 columns

#Checking the names of the columns in the dataset  
vert$name

#Checking the column datatypes  
sapply(vert, class)

#Summary statistics   
summary(vert)

#Information about the dataset  
glimpse(vert)

# Checking the column names  
names(vert)

DATA CLEANING

#Checking for missing values in the dataset  
is.na(vert)

The dataset has no missing values

#Checking for the total number of missing values in the columns as confirmation  
colSums(is.na(vert))

This confirms that the dataset does not have missing values

#Checking for duplicates in the dataset  
duplicated\_rows <- vert[duplicated(vert),]  
duplicated\_rows

The dataset has no duplicates

##TIDYING THE DATASET

#Checking for anomalies in the dataset  
anom <- lapply(vert,unique)  
anom

#Converting some columns into numerical datasset  
vert$Timestamp <- as.Date(vert$Timestamp)  
  
vert$"Clicked.on.Ad"<-as.factor(vert$"Clicked.on.Ad")  
vert$"Male" <- as.factor(vert$"Male")

##Checking For Outliers in the dataset

#Time spent on site  
boxplot(vert[,1])

The data contains outliers

#Age  
boxplot(vert[,2])

The column contains notable outliers suggesting the presence of extreme values in the dataset

#Daily internet usage  
boxplot(vert[,3])

The column has notable outliers signifying the presence of extreme values

#Male  
boxplot(vert[,7])

#Clicked on Ad  
boxplot(vert[,10])

The column cobntains notable outliers

####EXPLORATORY DATA ANALYSIS

#UNIVARIATE DATA ANALYSIS

#Summary statistics of the dataset  
   
summary(vert)

The mode, median, mean, maximum, minimum and range of the columns have been displayed above.

# Replacing spaces in the columns names with an underscoreto help calculate the frequency table  
names(vert) <- gsub(" ", "\_", names(vert))  
names(vert)

The names have been converted to an easy to use format

#Frequency Table  
library("plyr")

#Daily time spent on site  
count(vert, 'Daily.Time.Spent.on.Site')

#] "Area.Income"  
count(vert, 'Area.Income')

#Ad topic line  
count(vert, 'Ad.Topic.Line')

#Male  
count(vert, 'Male')

The males are less than other individuals

#Time stamp  
count(vert, 'Timestamp')

#Age  
count(vert, 'Age')

60 is the most common age

#City  
count(vert, 'City')

#Country  
count(vert, 'Country')

#Clicked on Ad  
count(vert, 'Clicked.on.Ad')

The ratio is the same

####Visualizations

#Male  
plot(factor(vert$"Male"), xlab ="Male", ylab ="Count", col= "green", main = "Male")

The number of participants who are not male is significantly higher.

#Age  
plot(factor(vert$"Age"), xlab ="Age", ylab ="Count", col= "red", main = "Age")

31 years is the most predominant age

##Histogram

####Histogram  
hist(vert$Area.Income)

The most predominant income is 60,000

#Daily Internet Usage  
hist(vert$Daily.Internet.Usage)

###BIVARIATE ANALYSIS

#Plotting a scatterplot between different variables  
plot(vert[, 1:4])

Their is no correlation between various variables

#Plotting the correlation matrix  
cor(vert[,1:4])

#Plotting a heat map   
heatmap(cor(vert[,1:4]))

Varibles that have a high correlation are: Daily Internet and Time spent on site, Area income and Usage, and Usage and Daily Time spent on site.

####Recommendations 1. The company should invest on more online adverts in areas that have a high income index. 2. Age of clients affected the their time in the internet, hence the company should evaluate some digital advertising techniques. 3. Areas that had more income had a high usage index and hence the compony is advised to explore that market.

###Conclusion The company should collect more data to support the change on digital marketing. #

####SUpervised Models  
  
The following models will be used to implement the target variable  
Regression, Decsion Tress and KNN.  
```{r}

#Loading the libraries needed  
library(party)  
library(rpart)

```{r}  
# Converting the target as a factor  
  
vert$Clicked.on.Ad = factor(vert$Clicked.on.Ad, levels = c(0,1))

# checking the variable datatypes  
  
sapply(vert, class)

#Viewing the data we will use  
head(vert)

# Specifying the target and predictor variables  
  
h <- rpart(Clicked.on.Ad ~ . ,   
 data = vert,   
 method = "class")

# Plotting the decision tree model  
  
rpart.plot(h)

# Using the above output to make a prediction  
pred <- predict(h, vert, type ="class")  
table(pred, vert$Clicked.on.Ad)

The diagonals represent the values used to predict customer ads.

#Predicting the models accuracy  
mean(vert$Clicked.on.Ad == pred)  
```{r}

The code predicted a good outcome. Their is need to challenge the model for valid results