Week 13 Unsupervised Learning IP 1

Brian Onchweri

5/30/2022

PREDICTING CUSTOMER PROPENSITY OF CLICKING AN AD USING UNSUPERVISED LEARNING MODEL.

Metric for Success

When i accurately determine which customer is likely to click an Ad

#Modelling

##Loading Libraries

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(data.table)
```

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

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

```r
library(tidyverse)
```

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

## v tibble  3.1.7     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x data.table::between() masks dplyr::between()
## x tidyr::extract()      masks magrittr::extract()
## x dplyr::filter()       masks stats::filter()
## x data.table::first()   masks dplyr::first()
## x dplyr::lag()          masks stats::lag()
## x data.table::last()    masks dplyr::last()
## x purrr::lift()         masks caret::lift()
## x purrr::set_names()    masks magrittr::set_names()
## x purrr::transpose()    masks data.table::transpose()
```

```r
library(rpart)
library(class)
require(class)
```

#Reading our dataset

```r
df<-read.csv('http://bit.ly/IPAdvertisingData')
head(df)
```

```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 1                    68.95  35    61833.90               256.09
## 2                    80.23  31    68441.85               193.77
## 3                    69.47  26    59785.94               236.50
## 4                    74.15  29    54806.18               245.89
## 5                    68.37  35    73889.99               225.58
## 6                    59.99  23    59761.56               226.74
##                          Ad.Topic.Line           City Male    Country
## 1    Cloned 5thgeneration orchestration    Wrightburgh    0    Tunisia
## 2    Monitored national standardization      West Jodi    1      Nauru
## 3      Organic bottom-line service-desk       Davidton    0 San Marino
## 4 Triple-buffered reciprocal time-frame West Terrifurt    1      Italy
## 5         Robust logistical utilization   South Manuel    0    Iceland
## 6       Sharable client-driven software      Jamieberg    1     Norway
##             Timestamp Clicked.on.Ad
## 1 2016-03-27 00:53:11             0
## 2 2016-04-04 01:39:02             0
## 3 2016-03-13 20:35:42             0
## 4 2016-01-10 02:31:19             0
## 5 2016-06-03 03:36:18             0
## 6 2016-05-19 14:30:17             0
```

```r
df<-data.table(df)
```

```r
# Changing the column names to lower case
names(df) <- tolower(names(df))
names(df)
```

```
##  [1] "daily.time.spent.on.site" "age"
##  [3] "area.income"              "daily.internet.usage"
##  [5] "ad.topic.line"            "city"
##  [7] "male"                     "country"
##  [9] "timestamp"                "clicked.on.ad"
```

```r
df$clicked.on.ad <- as.factor(df$clicked.on.ad)
```

```r
df$clicked.on.ad <- as.numeric(df$clicked.on.ad)
```

```r
head(df)
```

```
##    daily.time.spent.on.site age area.income daily.internet.usage
## 1:                    68.95  35    61833.90               256.09
## 2:                    80.23  31    68441.85               193.77
## 3:                    69.47  26    59785.94               236.50
## 4:                    74.15  29    54806.18               245.89
## 5:                    68.37  35    73889.99               225.58
## 6:                    59.99  23    59761.56               226.74
##                            ad.topic.line           city male    country
## 1:    Cloned 5thgeneration orchestration    Wrightburgh    0    Tunisia
## 2:    Monitored national standardization      West Jodi    1      Nauru
## 3:      Organic bottom-line service-desk       Davidton    0 San Marino
## 4: Triple-buffered reciprocal time-frame West Terrifurt    1      Italy
## 5:         Robust logistical utilization   South Manuel    0    Iceland
## 6:       Sharable client-driven software      Jamieberg    1     Norway
##              timestamp clicked.on.ad
## 1: 2016-03-27 00:53:11             1
## 2: 2016-04-04 01:39:02             1
## 3: 2016-03-13 20:35:42             1
## 4: 2016-01-10 02:31:19             1
## 5: 2016-06-03 03:36:18             1
## 6: 2016-05-19 14:30:17             1
```

```r
df1 <- select(df, c(1,2,3,4,7,10))
#df1 <- select(df1, -c(7,8))
head(df1)
```

```
##    daily.time.spent.on.site age area.income daily.internet.usage male
## 1:                    68.95  35    61833.90               256.09    0
## 2:                    80.23  31    68441.85               193.77    1
## 3:                    69.47  26    59785.94               236.50    0
## 4:                    74.15  29    54806.18               245.89    1
## 5:                    68.37  35    73889.99               225.58    0
## 6:                    59.99  23    59761.56               226.74    1
```

```
##    clicked.on.ad
## 1:             1
## 2:             1
## 3:             1
## 4:             1
## 5:             1
## 6:             1
```

```
length(df1$clicked.on.ad)
```

```
## [1] 1000
```

```
length(df1$area.income)
```

```
## [1] 1000
```

```
#Create an index for data partitioning

set.seed(7)

index<- createDataPartition(df1$clicked.on.ad,p = 0.8 ,list = FALSE)
```

```
#Using the indexes to split data into test and train set
df.train <- df1[index, ]
df.test <- df1[-index, ]
```

#Decision Trees

```
#Fitting in the decision tree
TreeFit <- rpart(clicked.on.ad ~ ., data = df.train ,method = "class")

#Factor the Clicked.on.Ad vector in the test dataset
df.test$clicked.on.ad <- factor(df.test$clicked.on.ad)

#Using model to predict
TreePredict <- predict(TreeFit, newdata = df.test, type = "class")
confusionMatrix(TreePredict, df.test$clicked.on.ad)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 97  6
##          2  3 94
##
##                Accuracy : 0.955
##                  95% CI : (0.9163, 0.9792)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.91
```

```
##
##  Mcnemar's Test P-Value : 0.505
##
##             Sensitivity : 0.9700
##             Specificity : 0.9400
##          Pos Pred Value : 0.9417
##          Neg Pred Value : 0.9691
##              Prevalence : 0.5000
##          Detection Rate : 0.4850
##    Detection Prevalence : 0.5150
##       Balanced Accuracy : 0.9550
##
##        'Positive' Class : 1
##
```

#KNN

```r
#Fitting model to training dataset
#Also we scale and center our data
knnModel <- train(as.factor(clicked.on.ad) ~ ., data =df.train, method = "knn", preProcess = c("center"
```

#Making Predictions

```r
#Using the model to predict
knnPredict <- predict(knnModel, newdata = df.test)
```

```r
#Printing out the confusion matrix and statistics
confusionMatrix(knnPredict, df.test$clicked.on.ad)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 100   6
##          2   0  94
##
##                Accuracy : 0.97
##                  95% CI : (0.9358, 0.9889)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.94
##
##  Mcnemar's Test P-Value : 0.04123
##
##             Sensitivity : 1.0000
##             Specificity : 0.9400
##          Pos Pred Value : 0.9434
##          Neg Pred Value : 1.0000
##              Prevalence : 0.5000
##          Detection Rate : 0.5000
##    Detection Prevalence : 0.5300
##       Balanced Accuracy : 0.9700
```

```
##
##        'Positive' Class : 1
##
```

#WE CAN SEE FROM ABOVE THE KNN WAS SLIGHTLLY MORE ACCURATE THAN DECISION TREE BUT GENERALLY THEY WERE BOTH VERY ACCURATE