R Project - Identifying individuals most likely to click an ad

Geoffrey Chege

2022-03-28

## 1. Introduction

### 1.1 Defining the question

* Determine which individuals are most likely to click on an ad using supervised learning prediction models.

### 1.2 The Context

* A Kenyan entrepreneur has created an online cryptography course and would want to advertise it on her blog.
* She currently targets audiences originating from various countries.
* In the past, she ran ads to advertise a related course on the same blog and collected data in the process.
* She would now like to employ my services as a Data Science Consultant to help her identify which individuals are most likely to click on her ads.

### 1.3 Metric for success

* Accuracy score of 85% and above.

### 1.4 Experimental Design Taken

* Installing packages and loading libraries needed
* Loading the data
* Data Cleaning
* Exploratory Data Analysis:
  + Univariate Analysis
  + Bivariate Analysis
* Modelling
* Predictions and evaluation of the model
* Conclusion

### 1.5 Appropriateness of the available data

* The columns in the dataset include:
  + Daily\_Time\_Spent\_on\_Site
  + Age
  + Area\_Income
  + Daily\_Internet\_Usage
  + Ad\_Topic\_Line
  + City
  + Male
  + Country
  + Timestamp
  + Clicked\_on\_Ad

## 2. Installing and loading Necessary Packages

## 3. Loading the Data

ad <- read.csv("C:/Users/user/Downloads/advertising.csv")  
head(ad)

## 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

## 4. Data Cleaning

### 4.1 Checking the attribute types

## Daily.Time.Spent.on.Site Age Area.Income   
## "numeric" "integer" "numeric"   
## Daily.Internet.Usage Ad.Topic.Line City   
## "numeric" "character" "character"   
## Male Country Timestamp   
## "integer" "character" "character"   
## Clicked.on.Ad   
## "integer"

### 4.2 converting time variable from character to date and time (POSIXct) format

ad$Timestamp <- as.POSIXct(ad$Timestamp, "%Y-%m-%d %H:%M:%S",tz = "GMT")

### 4.3 Checking for duplicates

duplicates <- ad[duplicated(ad),]  
  
duplicates

## [1] Daily.Time.Spent.on.Site Age Area.Income   
## [4] Daily.Internet.Usage Ad.Topic.Line City   
## [7] Male Country Timestamp   
## [10] Clicked.on.Ad   
## <0 rows> (or 0-length row.names)

There are no duplicates in the dataset

### 4.4 checking for null values

colSums(is.na(ad))

## Daily.Time.Spent.on.Site Age Area.Income   
## 0 0 0   
## Daily.Internet.Usage Ad.Topic.Line City   
## 0 0 0   
## Male Country Timestamp   
## 0 0 0   
## Clicked.on.Ad   
## 0

There are no null values in the dataset

### 4.5 checking column names

names(ad)

## [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"

Replacing the periods “.” with underscores “\_”

names(ad) <- gsub("[.]", "\_", names(ad))

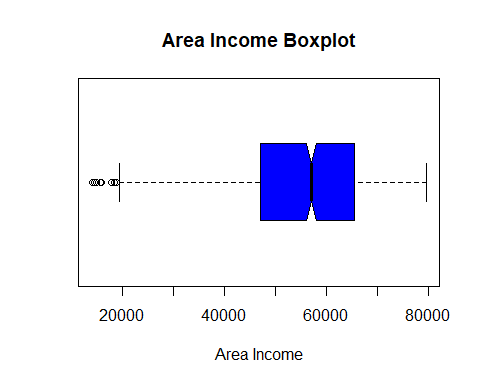
names(ad)

## [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"

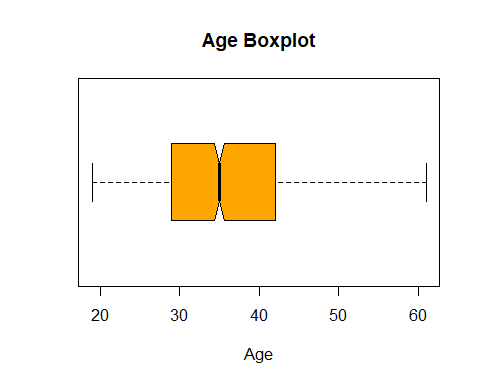
### 4.6 Outliers

I will use boxplots to check for outliers.

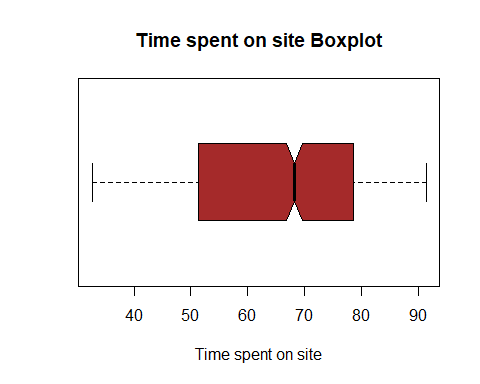
#### Boxplot for “Area\_Income”



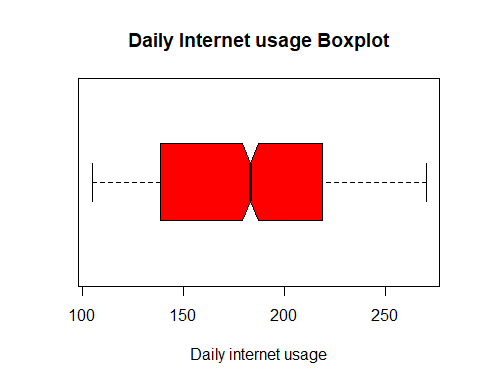
#### Boxplot for “Age”



#### Boxplot for “Daily\_Time\_Spent\_on\_Site”



#### Boxplot for “Daily\_Internet\_Usage”



## 5. Exploratory Data Analysis

### 5.1 Univariate Analysis

Summary statistics of the dataset

summary(ad)

## Daily\_Time\_Spent\_on\_Site Age Area\_Income Daily\_Internet\_Usage  
## Min. :32.60 Min. :19.00 Min. :13996 Min. :104.8   
## 1st Qu.:51.36 1st Qu.:29.00 1st Qu.:47032 1st Qu.:138.8   
## Median :68.22 Median :35.00 Median :57012 Median :183.1   
## Mean :65.00 Mean :36.01 Mean :55000 Mean :180.0   
## 3rd Qu.:78.55 3rd Qu.:42.00 3rd Qu.:65471 3rd Qu.:218.8   
## Max. :91.43 Max. :61.00 Max. :79485 Max. :270.0   
## Ad\_Topic\_Line City Male Country   
## Length:1000 Length:1000 Min. :0.000 Length:1000   
## Class :character Class :character 1st Qu.:0.000 Class :character   
## Mode :character Mode :character Median :0.000 Mode :character   
## Mean :0.481   
## 3rd Qu.:1.000   
## Max. :1.000   
## Timestamp Clicked\_on\_Ad  
## Min. :2016-01-01 02:52:10 Min. :0.0   
## 1st Qu.:2016-02-18 02:55:42 1st Qu.:0.0   
## Median :2016-04-07 17:27:29 Median :0.5   
## Mean :2016-04-10 10:34:06 Mean :0.5   
## 3rd Qu.:2016-05-31 03:18:14 3rd Qu.:1.0   
## Max. :2016-07-24 00:22:16 Max. :1.0

From the summary statistics, the following can be concluded about these columns:

#### Daily\_Time\_Spent\_on\_Site:

mean: 65

median: 68.22

#### Age:

mean: 36.01

median: 35

#### Area Income:

mean: 55,000

median: 57,012

#### Daily\_Internet\_Usage:

mean: 180

median: 183.1

Using describe() to get range, skewness, kurtosis and standard deviation among others:

describe(ad)

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

## vars n mean sd median trimmed mad  
## Daily\_Time\_Spent\_on\_Site 1 1000 65.00 15.85 68.22 65.74 17.92  
## Age 2 1000 36.01 8.79 35.00 35.51 8.90  
## Area\_Income 3 1000 55000.00 13414.63 57012.30 56038.94 13316.62  
## Daily\_Internet\_Usage 4 1000 180.00 43.90 183.13 179.99 58.61  
## Ad\_Topic\_Line\* 5 1000 500.50 288.82 500.50 500.50 370.65  
## City\* 6 1000 487.32 279.31 485.50 487.51 356.57  
## Male 7 1000 0.48 0.50 0.00 0.48 0.00  
## Country\* 8 1000 116.41 69.94 114.50 115.82 89.70  
## Timestamp 9 1000 NaN NA NA NaN NA  
## Clicked\_on\_Ad 10 1000 0.50 0.50 0.50 0.50 0.74  
## min max range skew kurtosis se  
## Daily\_Time\_Spent\_on\_Site 32.60 91.43 58.83 -0.37 -1.10 0.50  
## Age 19.00 61.00 42.00 0.48 -0.41 0.28  
## Area\_Income 13996.50 79484.80 65488.30 -0.65 -0.11 424.21  
## Daily\_Internet\_Usage 104.78 269.96 165.18 -0.03 -1.28 1.39  
## Ad\_Topic\_Line\* 1.00 1000.00 999.00 0.00 -1.20 9.13  
## City\* 1.00 969.00 968.00 0.00 -1.19 8.83  
## Male 0.00 1.00 1.00 0.08 -2.00 0.02  
## Country\* 1.00 237.00 236.00 0.08 -1.23 2.21  
## Timestamp Inf -Inf -Inf NA NA NA  
## Clicked\_on\_Ad 0.00 1.00 1.00 0.00 -2.00 0.02

#### Mode

A function to determine the mode:

mode <- function(v){  
 uniq <- unique(v)  
 uniq[which.max(tabulate(match(v,uniq)))]  
}

The most recurrent Ad Topic Line:

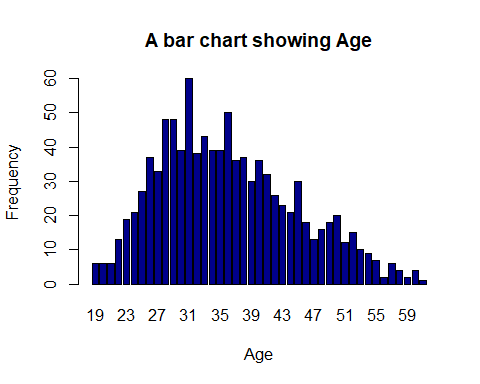
## [1] "Cloned 5thgeneration orchestration"

The most recurrent City:

## [1] "Lisamouth"

The most recurrent Country:

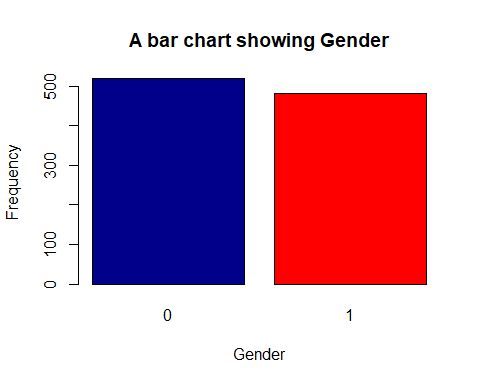
## [1] "Czech Republic"

Checking the modal age using a barplot: 

From the plot, the modal age is 31.

Checking the distribution in terms of gender where 1 is Male and 0 is Female:

## gender  
## 0 1   
## 519 481

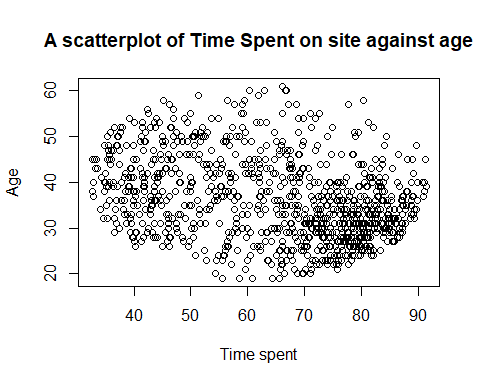


From this, there are More women than men, making female the modal gender.

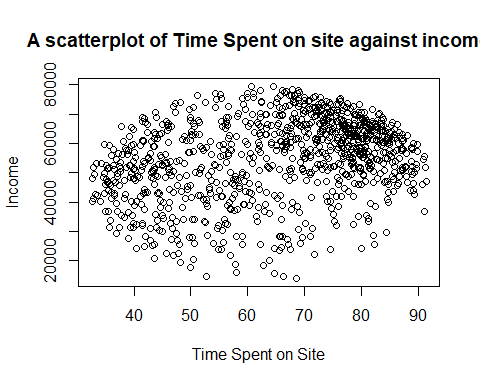
### 5.2 Bivariate Analysis

#### Scatterplots

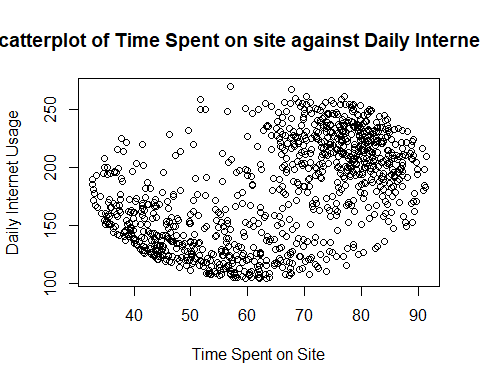
# scatterplot  
plot((ad$Daily\_Time\_Spent\_on\_Site), (ad$Age),   
 main = "A scatterplot of Time Spent on site against age",  
 xlab = 'Time spent',   
 ylab = 'Age')



# scatterplot of Time on site vs income  
plot((ad$Daily\_Time\_Spent\_on\_Site), (ad$Area\_Income),   
 main = "A scatterplot of Time Spent on site against income",  
 xlab = 'Time Spent on Site',   
 ylab = 'Income')

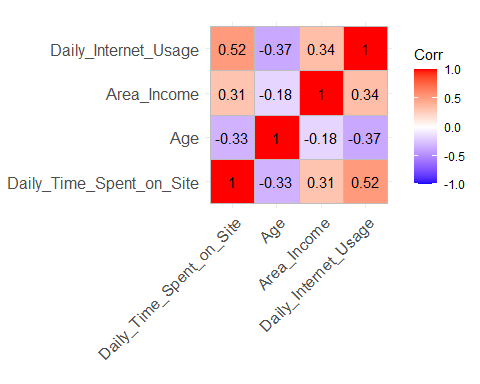


# scatterplot of Time on site vs Internet usage  
plot((ad$Daily\_Time\_Spent\_on\_Site), (ad$Daily\_Internet\_Usage),   
 main = "A scatterplot of Time Spent on site against Daily Internet Usage",  
 xlab = 'Time Spent on Site',   
 ylab = 'Daily Internet Usage')



#### Heatmap

# Heat map  
# Checking the relationship between the variables  
# Using Numeric variables only  
  
numeric\_tbl <- ad %>%  
 select\_if(is.numeric) %>%  
 select(Daily\_Time\_Spent\_on\_Site, Age, Area\_Income,Daily\_Internet\_Usage)  
  
# Calculate the correlations  
corr <- cor(numeric\_tbl, use = "complete.obs")  
  
ggcorrplot(round(corr, 2),   
 type = "full", lab = T)

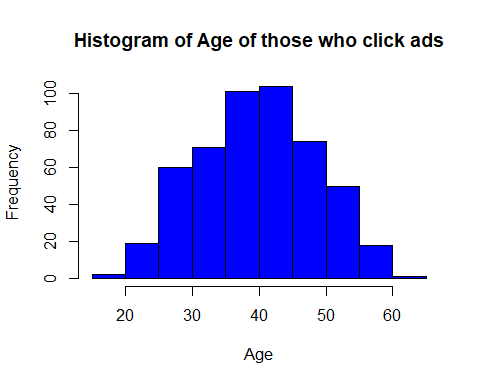
 #### Those who clicked on ads:

Analysis of people who click on the ads:

# Analysis of people who click on the ads  
ad\_click <- ad[which(ad$Clicked\_on\_Ad == 1),]

Most popular age group of people clicking on ads:

# Most popular age group of people clicking on ads  
hist((ad\_click$Age),   
 main = "Histogram of Age of those who click ads",  
 xlab = 'Age',   
 ylab = 'Frequency',  
 col = "blue")



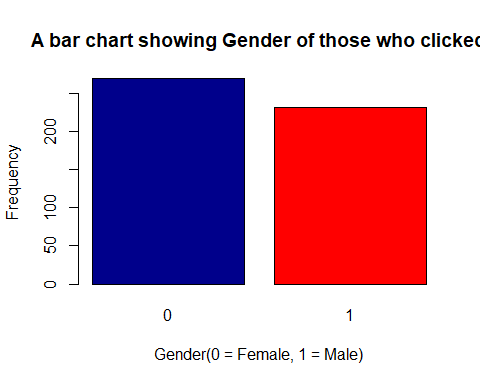
40 - 45 year olds click on the most ads

#### Plotting to visualize the gender distribution:

gender2 <- (ad\_click$Male)  
gender2.frequency <- table(gender2)  
gender2.frequency

## gender2  
## 0 1   
## 269 231

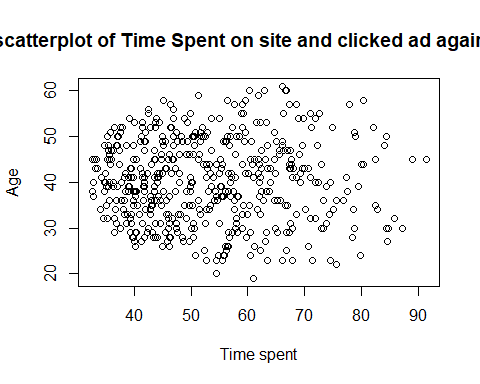
# plotting to visualize the gender distribution  
  
barplot(gender2.frequency,  
 main="A bar chart showing Gender of those who clicked",  
 xlab="Gender(0 = Female, 1 = Male)",  
 ylab = "Frequency",  
 col=c("darkblue","red"),  
 )



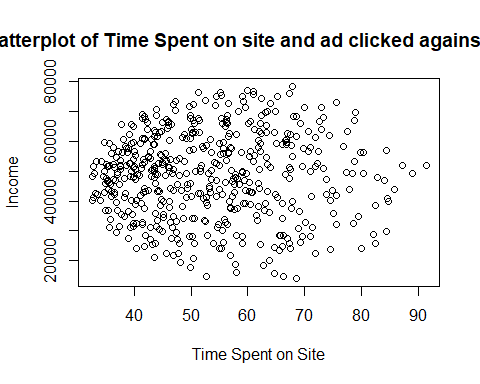
Females clicked more ads than males.

#### Scatterplots of those who clicked:

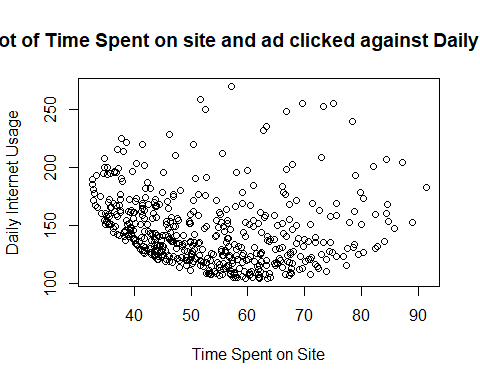
# scatterplot  
plot((ad\_click$Daily\_Time\_Spent\_on\_Site), (ad\_click$Age),   
 main = "A scatterplot of Time Spent on site and clicked ad against age",  
 xlab = 'Time spent',   
 ylab = 'Age')



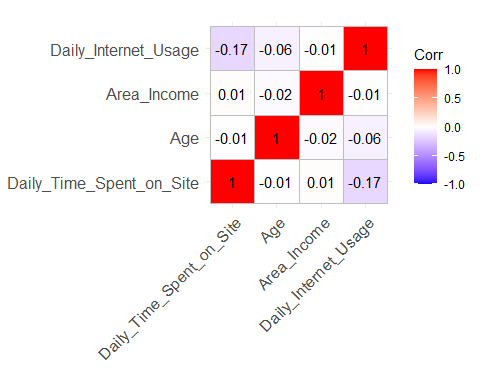
# scatterplot of Time on site vs income  
plot((ad\_click$Daily\_Time\_Spent\_on\_Site), (ad\_click$Area\_Income),   
 main = "A scatterplot of Time Spent on site and ad clicked against income",  
 xlab = 'Time Spent on Site',   
 ylab = 'Income')



# scatterplot of Time on site vs Internet usage  
plot((ad\_click$Daily\_Time\_Spent\_on\_Site), (ad\_click$Daily\_Internet\_Usage),   
 main = "A scatterplot of Time Spent on site and ad clicked against Daily Internet Usage",  
 xlab = 'Time Spent on Site',   
 ylab = 'Daily Internet Usage')



# Heat map  
# Checking the relationship between the variables  
# Using Numeric variables only  
  
numeric\_tbl <- ad\_click %>%  
 select\_if(is.numeric) %>%  
 select(Daily\_Time\_Spent\_on\_Site, Age, Area\_Income,Daily\_Internet\_Usage)  
  
# Calculate the correlations  
corr <- cor(numeric\_tbl, use = "complete.obs")  
  
ggcorrplot(round(corr, 2),   
 type = "full", lab = T)



The country with the most ad clicks:

mode(ad\_click$Country)

## [1] "Australia"

The income that clicks most:

mode(ad\_click$Area\_Income)

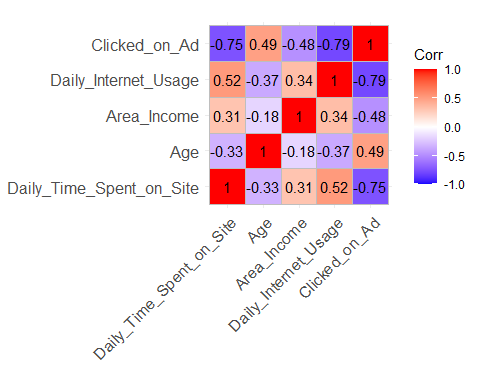
## [1] 24593.33

Ad title that garners most clicks:

## [1] "Reactive local challenge"

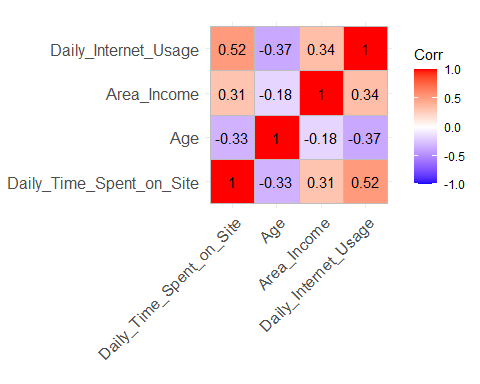
## 6. Modelling

# Heat map  
# Checking the relationship between the variables  
# Using Numeric variables only  
  
numeric\_tbl2 <- ad %>%  
 select\_if(is.numeric) %>%  
 select(Daily\_Time\_Spent\_on\_Site, Age, Area\_Income,Daily\_Internet\_Usage, Clicked\_on\_Ad)  
  
# Calculate the correlations  
corr <- cor(numeric\_tbl2, use = "complete.obs")  
  
ggcorrplot(round(corr, 2),   
 type = "full", lab = T)



Clicked\_on\_Ad has high correlation. It is the target variable so I will not include it in the correlation.

# Heat map  
# Checking the relationship between the variables  
# Using Numeric variables only  
  
numeric\_tbl3 <- ad %>%  
 select\_if(is.numeric) %>%  
 select(Daily\_Time\_Spent\_on\_Site, Age, Area\_Income,Daily\_Internet\_Usage)  
  
# Calculate the correlations  
corr <- cor(numeric\_tbl3, use = "complete.obs")  
  
ggcorrplot(round(corr, 2),   
 type = "full", lab = T)

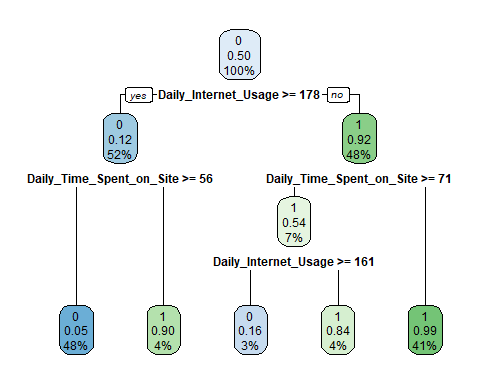


There are no highly correlated numeric independent variables, so I will use them all in analysis.

### Decision Tree Classifier

# Specifying target and predictor variables  
  
m <- rpart(Clicked\_on\_Ad ~ . ,  
 data = numeric\_tbl2,  
 method = "class")

# Plotting model  
  
rpart.plot(m)



# Making predictions  
# Printing the confusion matrix  
  
p <- predict(m, numeric\_tbl2, type ="class")  
table(p, numeric\_tbl2$Clicked\_on\_Ad)

##   
## p 0 1  
## 0 485 28  
## 1 15 472

# Printing the Accuracy  
  
(mean(numeric\_tbl2$Clicked\_on\_Ad == p))\*100

## [1] 95.7

* The model has an accuracy of 95.7%
* This is a good model for making predictions

## 7. Conclusion

Decision Tree gives an accuracy of 95.7%