Real-Time Prediciton of Online Purchase Behavior

Harini

03/12/2020

ACKNOWLEDGEMENT

I wish to express my sense of gratitude and sincere thanks to Dr.Rafael Irizarry, Professor of Biostatistics at Harvard Chan School of Public Health, for his kind support. I also extend my sincere thanks to the Teaching Assistants who gave great inputs and for their valuable suggestion throughout the Data Science course.

INTRODUCTION

The Project is related to the choose your own project of the HarvardX:PH125:9x Data science Capstone. Now-a-days, due to technological advancement more customer choose Internet platform to buy their products as it is easy and convenient. It has become very essential to know the customer needs for any online merchants to sustain in such competitive market. The records of the consumer operations and consumer behavior data, make it possible to predict customers buying preferences. This empirical study investigates the contribution of different types of predictors to the purchasing behaviour at an online store.

PROBLEM DEFINITION

Accurate prediction of shopping channel preferences has become an important issue for retailers seeking to maximize customer loyalty. We evaluate the predictive accuracy of an unbalanced classification of consumer online shopping behaviour using Clustering and Classification algorithms. The main objective of this project is to find the key metrics which contributes the most to predict online purchase behavior. This project also give some suggestions to improve the performance of e-shopping platform. The data is collected from the UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/machine-learning-databases/00468/online_shoppers_intention.csv. The dataset has 12,330, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.

DATA INGESTION

The dataset is in the .csv format. It consist of 10 numerical and 8 categorical variables.The numerical variables of the dataset were normalized for clustering and classification methods. The 70% of the data were used to train the dataset and our models were evaluated on the remaining 10% of Validation set.

The data frame has 18 variables. The variables Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration tells about the e-merchant website pages. The website visited by the shopper in specific session and their total time spent in each of these pages. These records were collected from the Uniform Resource Locator information of the pages visited by the consumer. The data also has Google Analytics metrics such as BounceRates, ExitRates, PageValues. Bounce rate refers to the first page a visitor enters, and exit rate refers to the last page they visits before they leaves. Bounce rate is the average number of bounces across all the pages divided by the total number of visits across all of those pages within the same period. This can tell that the searching result of consumer does not match their intent well. The average bounce rate is 58.18 percentage for B2C businesses. The last page from the shoppers journey of sites is considered an exit page, and it will contribute to determining Exit Rate. The exit rate can be high if the shoppers found the information they needed, and then left the page. Page Value is the average value for a page that a shopper visited before landing on our page or completing an E-commerce transaction (or both). Special Day represents any festival season where we would have more transactions. The dataset also has different information about the shoppers operating system, browser, region, traffic and visitor type. It also has month of the shoppers visit and a Boolean value indicating whether its a weekend or not. Our target variable is Revenue that says about the customer has purchased on our website or not. Sparkling the curiosity of customer is very essential and making them want to explore instead of leaving website will do wonders in an e-business! And hence these variables are very important to understand. The preview of structure of the data is given below. There are no missing values in the dataset.

```
str(data)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : chr  "Feb" "Feb" "Feb" "Feb" ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : chr  "Returning_Visitor" "Returning_Visitor" "Returning_Visitor" "Returni
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
head(data)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                       0             0                      0
## 4              0                       0             0                      0
## 5              0                       0             0                      0
## 6              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1                0.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
```

```
## 1          0   Feb                  1       1      1            1
## 2          0   Feb                  2       2      1            2
## 3          0   Feb                  4       1      9            3
## 4          0   Feb                  3       2      2            4
## 5          0   Feb                  3       3      1            4
## 6          0   Feb                  2       2      1            3
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   FALSE   FALSE
## 2 Returning_Visitor   FALSE   FALSE
## 3 Returning_Visitor   FALSE   FALSE
## 4 Returning_Visitor   FALSE   FALSE
## 5 Returning_Visitor    TRUE   FALSE
## 6 Returning_Visitor   FALSE   FALSE
```

```r
summary(data) #summary statistics
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##   BounceRates         ExitRates          PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##     Month           OperatingSystems    Browser           Region
##  Length:12330       Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.124    Mean   : 2.357   Mean   :3.147
##                     3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##                     Max.   :8.000    Max.   :13.000   Max.   :9.000
##   TrafficType    VisitorType          Weekend          Revenue
##  Min.   : 1.00   Length:12330       Mode :logical   Mode :logical
##  1st Qu.: 2.00   Class :character   FALSE:9462      FALSE:10422
##  Median : 2.00   Mode  :character   TRUE :2868      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
```

```r
##Missing value analysis
colSums(is.na(data))
```

```
##          Administrative Administrative_Duration           Informational
##                       0                       0                       0
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                       0                       0                       0
##             BounceRates               ExitRates              PageValues
##                       0                       0                       0
##              SpecialDay                   Month        OperatingSystems
##                       0                       0                       0
##                 Browser                  Region             TrafficType
##                       0                       0                       0
##             VisitorType                 Weekend                 Revenue
##                       0                       0                       0
```

DATA PREPROCESSING

The structure of the variables were altered according to categorical and numerical basis. Now, the categorical variables were converted into ordered factor variables and numerically encoded. The new dataset look like:

```
## 'data.frame':    12330 obs. of  20 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Ord.factor w/ 10 levels "Feb"<"Mar"<"May"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ OperatingSystems       : Factor w/ 8 levels "1","2","3","4",..: 1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : Factor w/ 13 levels "1","2","3","4",..: 1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : Factor w/ 9 levels "1","2","3","4",..: 1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ Revenue                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weekend_01             : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue_01             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

EXPLORATORY DATA ANALYSIS

The summary statistics of the dataset is given below

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :    0.00        Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:    0.00        1st Qu.: 0.0000
##  Median : 1.000   Median :    7.50        Median : 0.0000
```

```
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
```

Lets us explore all variables. The distribution of Revenue tells us that the Revenue turned out is 15 Percent.

```
##
##     0     1
## 10422  1908
```

The Distribution of Weekend is

```
##
##    0    1
## 9462 2868
```

The Distribution of Visitor Type is

```
##
##       New_Visitor             Other Returning_Visitor
##              1694                85             10551
```

The Distribution of Traffic Type is

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2451 3913 2052 1069  260  444   40  343   42  450  247    1  738   13   38    3
##   17   18   19   20
##    1   10   17  198
```

The Distribution of Region is

```
##
##    1    2    3    4    5    6    7    8    9
## 4780 1136 2403 1182  318  805  761  434  511
```

The Distribution of Browser is

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13
## 2462 7961  105  736  467  174   49  135    1  163    6   10   61
```

The Distribution of Operating Systems is

```
##
##    1    2    3    4    5    6    7    8
## 2585 6601 2555  478    6   19    7   79
```

The Distribution of month is

```
##
##  Feb  Mar  May June  Jul  Aug  Sep  Oct  Nov  Dec
##  184 1907 3364  288  432  433  448  549 2998 1727
```

The summary statistics of Administrative is

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.315   4.000  27.000
```

The summary statistics of Administrative_Duration is

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    7.50   80.82   93.26 3398.75
```

The summary statistics of Informational is

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.5036  0.0000 24.0000
```

The summary statistics of Informational_Duration is

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   34.47    0.00 2549.38
```

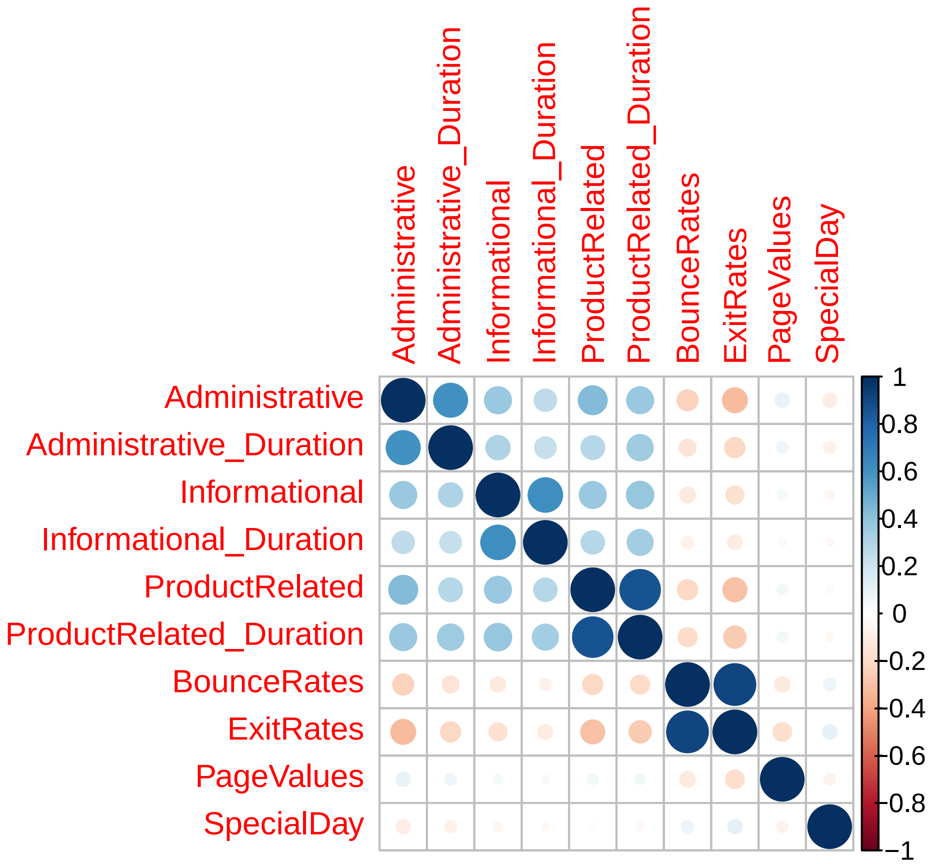
The summary statistics of Product_Related is

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   18.00   31.73   38.00  705.00
```

The summary statistics of Product_Related_Duration is

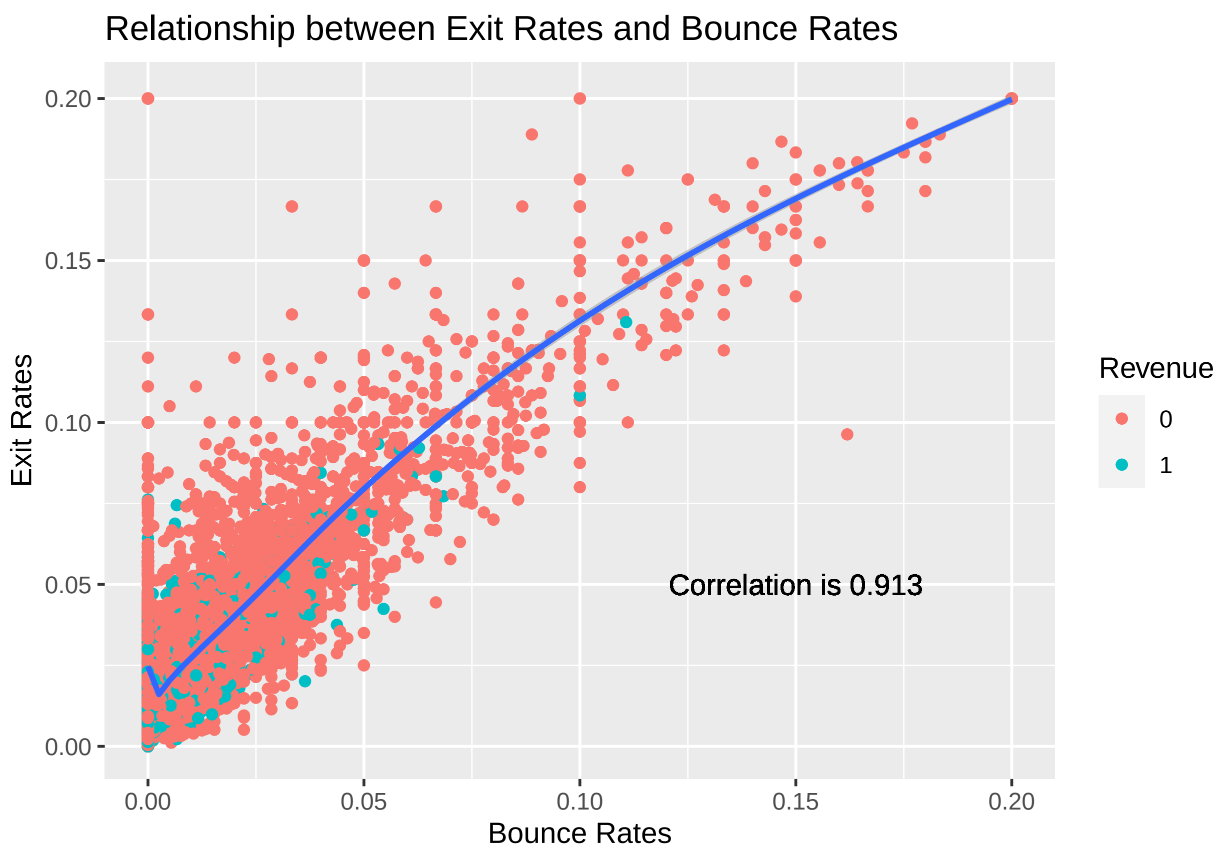
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   184.1   598.9  1194.8  1464.2 63973.5
```

Let us perform correlation analysis, which is used to quantify the association between two quantitative variables.

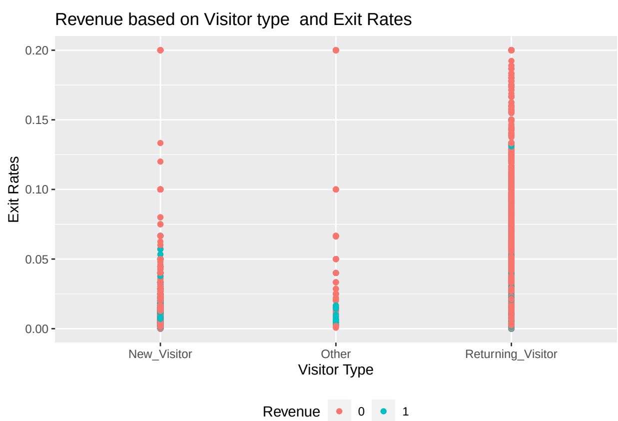


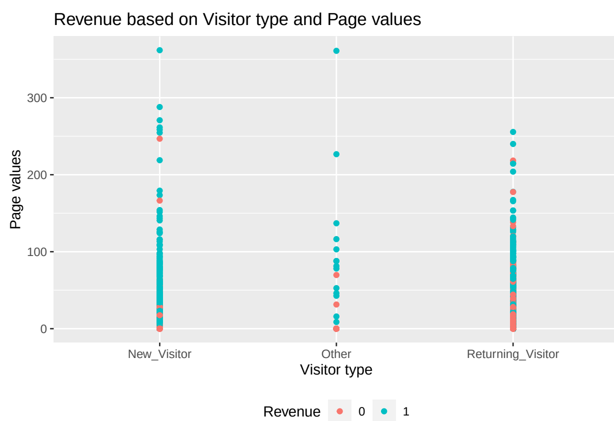
Let us plot the relationship between Bounce Rates and Exit Rates. It is evident from the plot, the shoppers who exit early are some of our potential customers. It is wise show some attractive pop ups like discount or huge offer when a customer attempt to leave the site.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

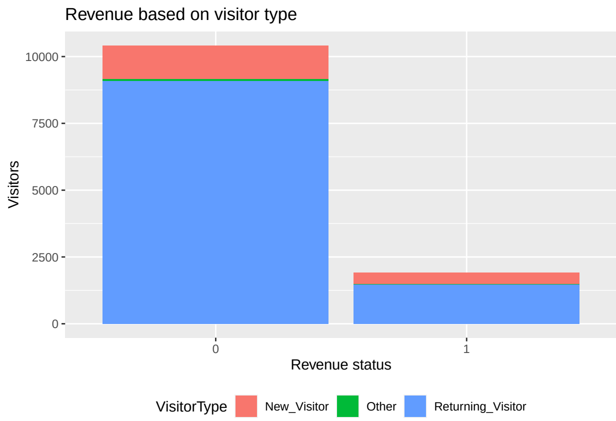


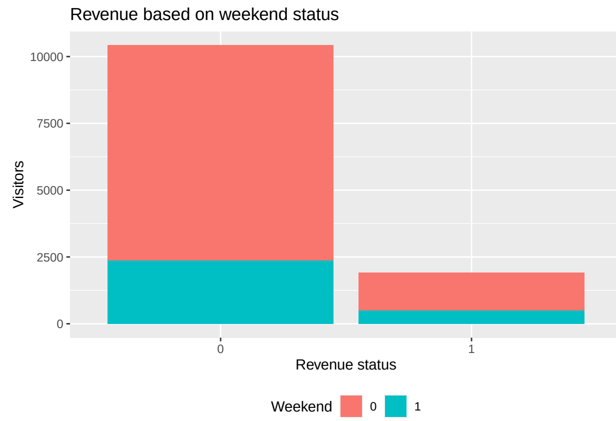
When we explore the relationship between visitor type-Exit Rate and visitor type-page values with respective to Revenue, the new visitor contributes more revenue than the returning visitor. Offering the reference coupons and giving discounts on it can bring new customers.



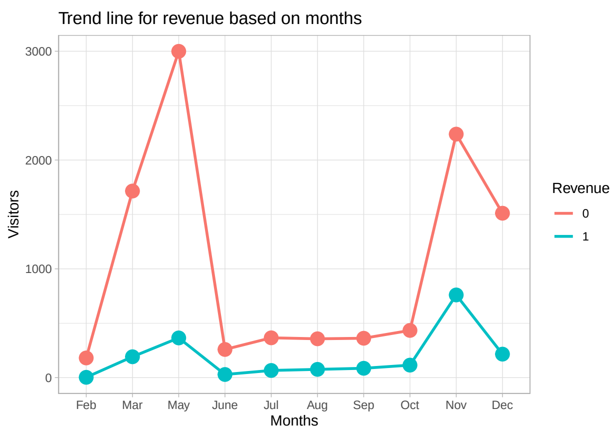


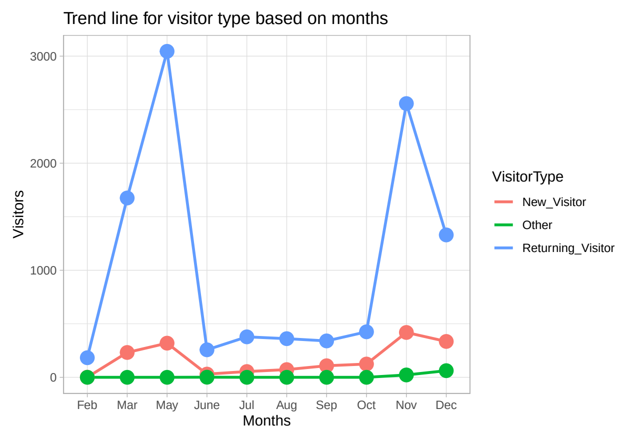
The conversion rate of potential customers is very important. Concentrating on new customers will significantly improve the sales and revenue growth. From the below plot, the purchase made during the weekday is higher than the weekends. Introducing weekends based promotional events may help the shoppers to engage during weekends.





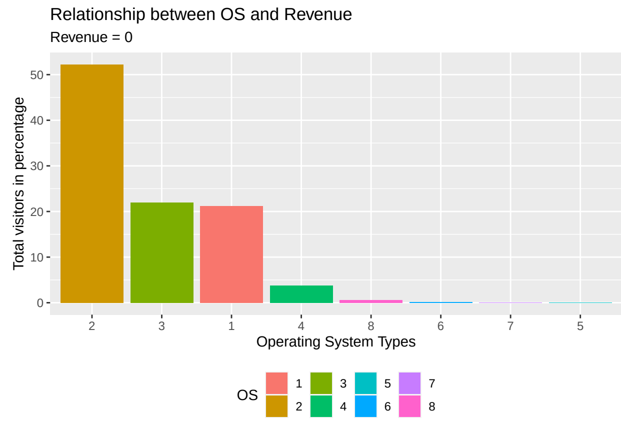
The below plot explain the seasonality revenue improvement. There seems to be many customers buy products during March to May and October to November. The plot also suggests that lot of customer are viewing the item but final transactions are made after adding into the cart. There may be hidden charges which may lead to loose the customers. Attractive offers and promotional events during festive season may engage more customers.

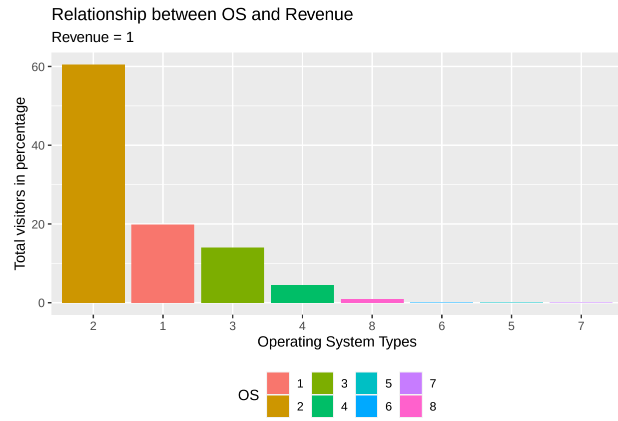




The operating systems of the user may also be considered as significant characteristics of predicting the shoppers. Most of our customer uses '2' OS type. Other OS are used by less customers. This could also mean many customer are not preferring to use the site in other sources.

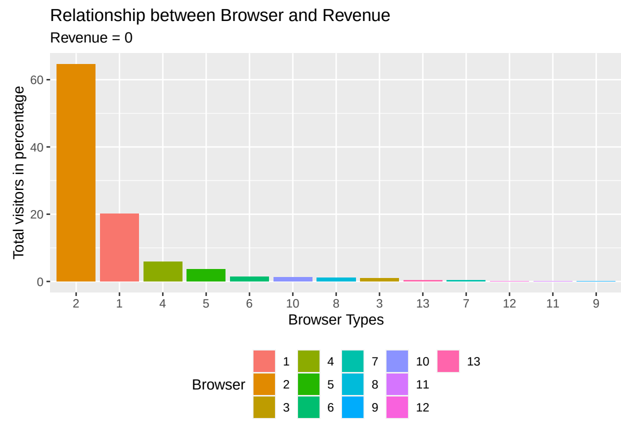
```
## 'data.frame':    16 obs. of  3 variables:
##  $ Var1: Factor w/ 8 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 1 2 ...
##  $ Var2: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Freq: int  2206 5446 2287 393 5 17 6 62 379 1155 ...
```

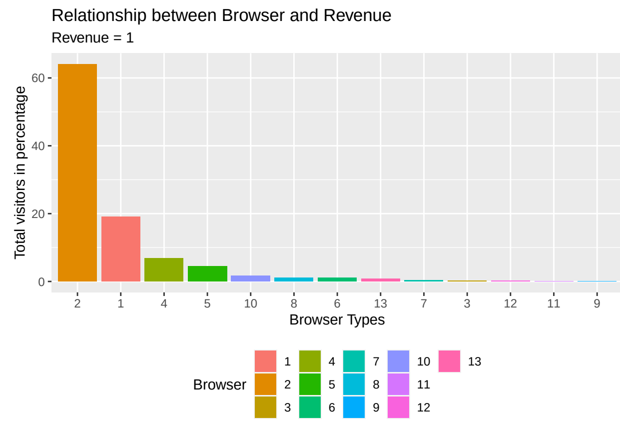




The relationship between Browser and Revenue states that the type '2' remains at the top. This may also suggest the website is not user friendly with other type of browsers. Web designers can concentrate on this for better improvement.

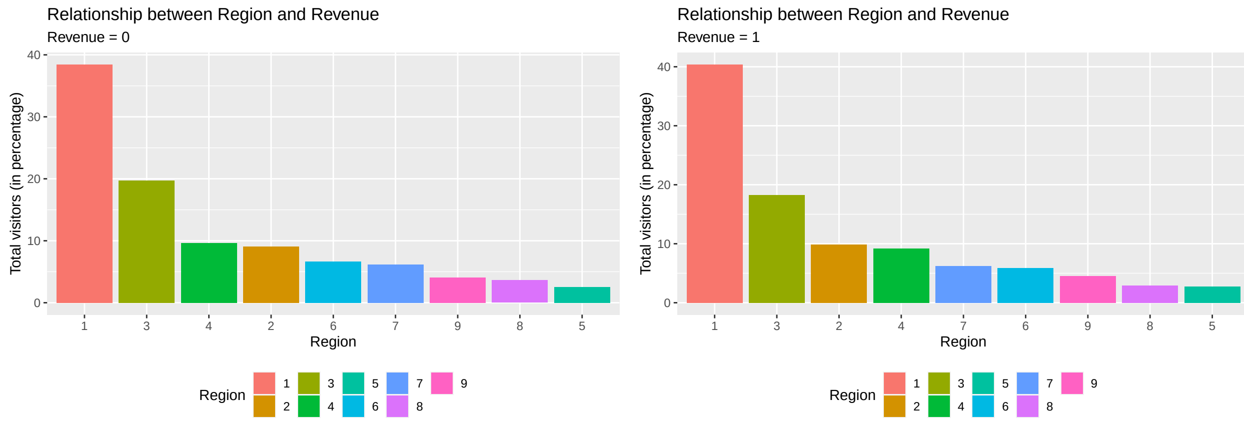
```
## 'data.frame':    26 obs. of  3 variables:
##  $ Var1: Factor w/ 13 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Var2: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Freq: int  2097 6738 100 606 381 154 43 114 1 131 ...
```





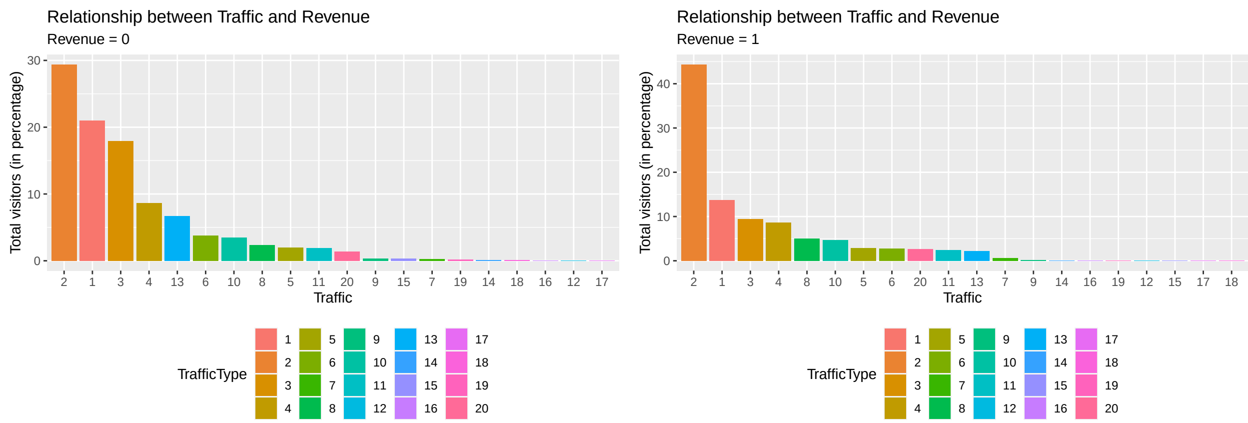
The relationship between Region and Revenue states that the most of our customers are from '1' and '3'. The marketing reach strategy can be helpful in these regions.

```
## 'data.frame':    18 obs. of  3 variables:
##  $ Var1: Factor w/ 9 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 1 ...
##  $ Var2: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Freq: int  4009 948 2054 1007 266 693 642 378 425 771 ...
```



The relationship plot between Traffic and Revenue states the type '2' traffic leads 'type1' and '3'.The Google SEO optimization can bring some improvement. Digital marketing in social media via ads can also bring significant customers.

```
## 'data.frame':    40 obs. of  3 variables:
##  $ Var1: Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Var2: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Freq: int  2189 3066 1872 904 204 391 28 248 38 360 ...
```



MODEL PREPARATION

In this project we used clustering and classification algorithms. And hence it is very essential to prepare our data for our models. Here we change all variable levels into factors with numeric levels. The distance between data points are important. Scaling the numeric data is very essential for certain machine learning models as we can maintain the same distribution of attributes. Then, removing the unwanted columns for evaluation.

```
## 'data.frame':    12330 obs. of  22 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
```

```
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Ord.factor w/ 10 levels "Feb"<"Mar"<"May"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ OperatingSystems       : Ord.factor w/ 8 levels "6"<"3"<"7"<"1"<..: 4 6 7 2 2 6 6 4 6 6 ...
##  $ Browser                : Ord.factor w/ 13 levels "9"<"3"<"6"<"7"<..: 5 6 5 6 2 6 9 6 6 9 ...
##  $ Region                 : Ord.factor w/ 9 levels "8"<"6"<"3"<"4"<..: 6 6 9 8 6 6 3 6 8 6 ...
##  $ TrafficType            : Ord.factor w/ 20 levels "12"<"15"<"17"<..: 9 16 7 11 11 7 7 15 7 16 ...
##  $ VisitorType            : Ord.factor w/ 3 levels "Returning_Visitor"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weekend                : num  1 1 1 1 2 1 1 2 1 1 ...
##  $ Revenue                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weekend_01             : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue_01             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Month_numeric          : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VisitorType_Numeric    : Ord.factor w/ 3 levels "1"<"2"<"3": 1 1 1 1 1 1 1 1 1 1 ...
```

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.Training set is a subset to train a model; Test set is a subset to test the trained model. Here we are splitting the data into 70 : 30 ratio for training and validation set.

```
#Splitting the data
#Splitting the data into 70:30 ratio
model_data <- data[-c(17,18,21,22)] # model_data for classification models
set.seed(777, sample.kind="Rounding")# if using R 3.5 or earlier, use ‘set.seed(1)‘
```

```
## Warning in set.seed(777, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
#Data Partition
test_index <- createDataPartition(model_data$Revenue, p = 0.7, list=FALSE)
#Training set
train_data <- model_data[test_index,]
#Test set
test_data <- model_data[-test_index,]
```

MODEL CREATION

The exploratory data analysis clearly says there is no clear distribution patterns among all attributes. Clustering can provide surprising insights into your data. Hence, we can use a clustering algorithm to classify each data point into a specific group.K-means is a very powerful method for finding a known number of clusters while considering the entire dataset. The structure of the data for clustering algorithm is

```
## 'data.frame':    12330 obs. of  17 variables:
##  $ Administrative         : num  0 0 0 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : num  0.00142 0.00284 0.00142 0.00284 0.01418 ...
##  $ ProductRelated_Duration: num  0.00 1.00e-03 0.00 4.17e-05 9.81e-03 ...
##  $ BounceRates            : num  1 0 1 0.25 0.1 ...
```

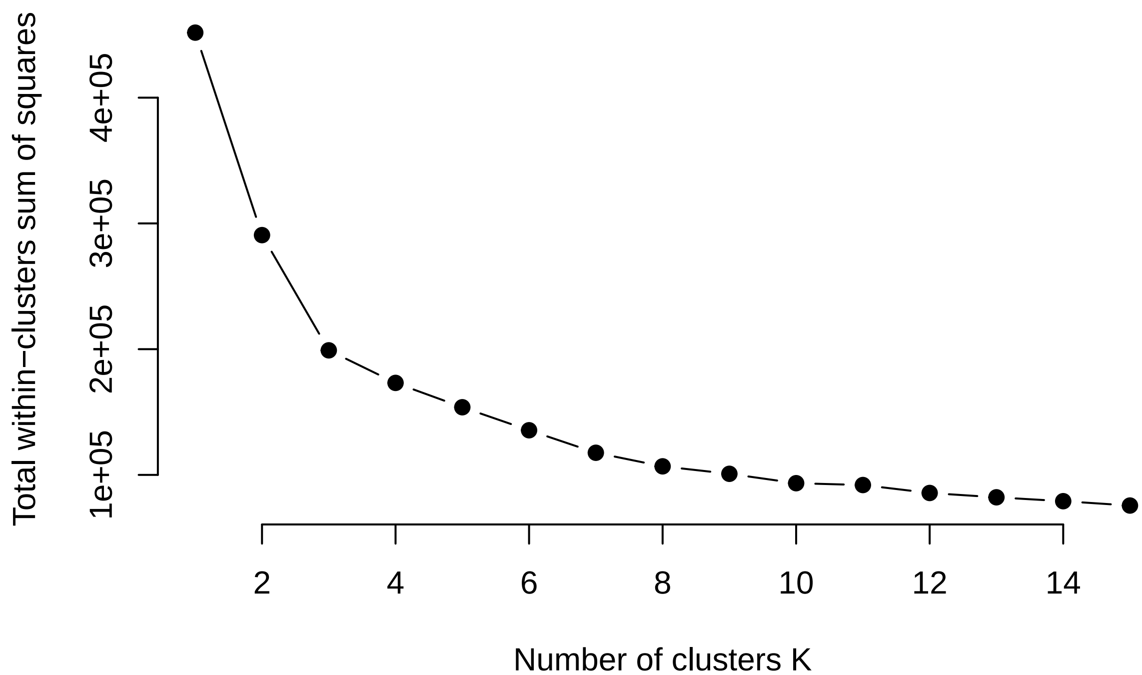
```
##  $ ExitRates              : num  1 0.5 1 0.7 0.25 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ OperatingSystems       : Ord.factor w/ 8 levels "6"<"3"<"7"<"1"<..: 4 6 7 2 2 6 6 4 6 6 ...
##  $ Browser                : Ord.factor w/ 13 levels "9"<"3"<"6"<"7"<..: 5 6 5 6 2 6 9 6 6 9 ...
##  $ Region                 : Ord.factor w/ 9 levels "8"<"6"<"3"<"4"<..: 6 6 9 8 6 6 3 6 8 6 ...
##  $ TrafficType            : Ord.factor w/ 20 levels "12"<"15"<"17"<..: 9 16 7 11 11 7 7 15 7 16 ...
##  $ Weekend                : num  1 1 1 1 2 1 1 2 1 1 ...
##  $ Month_numeric          : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VisitorType_Numeric    : Ord.factor w/ 3 levels "1"<"2"<"3": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  Administrative    Administrative_Duration Informational
##  Min.   :0.00000   Min.   :0.000000        Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000000        1st Qu.:0.00000
##  Median :0.03704   Median :0.002207        Median :0.00000
##  Mean   :0.08575   Mean   :0.023779        Mean   :0.02098
##  3rd Qu.:0.14815   3rd Qu.:0.027438        3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.000000        Max.   :1.00000
##
##  Informational_Duration ProductRelated     ProductRelated_Duration
##  Min.   :0.00000        Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000        1st Qu.:0.009929   1st Qu.:0.002878
##  Median :0.00000        Median :0.025532   Median :0.009362
##  Mean   :0.01352        Mean   :0.045009   Mean   :0.018676
##  3rd Qu.:0.00000        3rd Qu.:0.053901   3rd Qu.:0.022887
##  Max.   :1.00000        Max.   :1.000000   Max.   :1.000000
##
##   BounceRates        ExitRates         PageValues        SpecialDay
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.07143   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.01556   Median :0.12578   Median :0.00000   Median :0.00000
##  Mean   :0.11096   Mean   :0.21536   Mean   :0.01628   Mean   :0.06143
##  3rd Qu.:0.08406   3rd Qu.:0.25000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##  OperatingSystems    Browser         Region      TrafficType      Weekend
##  2      :6601     2      :7961   1      :4780   2      :3913   Min.   :1.000
##  1      :2585     1      :2462   3      :2403   1      :2451   1st Qu.:1.000
##  3      :2555     4      : 736   4      :1182   3      :2052   Median :1.000
##  4      : 478     5      : 467   2      :1136   4      :1069   Mean   :1.233
##  8      :  79     6      : 174   6      : 805   13     : 738   3rd Qu.:1.000
##  6      :  19     10     : 163   7      : 761   10     : 450   Max.   :2.000
##  (Other):  13     (Other): 367   (Other):1263   (Other):1657
##  Month_numeric  VisitorType_Numeric
##  3      :3364   1:10551
##  9      :2998   2:   85
##  2      :1907   3: 1694
##  10     :1727
##  8      : 549
##  7      : 448
##  (Other):1337
```

k-means consists of defining k clusters such that total within-cluster variation is minimum. To decide the number of optimal number of clusters we choose the Elbow Method. Calculate the Within-Cluster-Sum of

Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

```
##  [1] 451707.82 290686.60 199058.26 173159.69 153818.72 135500.22 117610.45
##  [8] 106806.17 100888.32  93415.87  91942.14  85626.85  82186.01  79105.02
## [15]  75649.31
```



The above plot above represents the variance within the clusters. The bend indicates that additional clusters beyond the fourth have little value. The R function kmeans() is used to compute k-means algorithm.

```r
str(k_means)
```

```
## List of 9
##  $ cluster     : int [1:12330] 1 1 1 1 1 1 1 1 1 1 ...
##  $ centers     : num [1:2, 1:17] 0.0741 0.1012 0.0208 0.0277 0.018 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:17] "Administrative" "Administrative_Duration" "Informational" "Informational_Dura
##  $ totss       : num 451708
##  $ withinss    : num [1:2] 266096 82890
##  $ tot.withinss: num 348986
##  $ betweenss   : num 102722
##  $ size        : int [1:2] 7029 5301
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```r
#size of cluster
k_means$size
```

```
## [1] 7029 5301
```

```r
#Means
k_means$centers
```

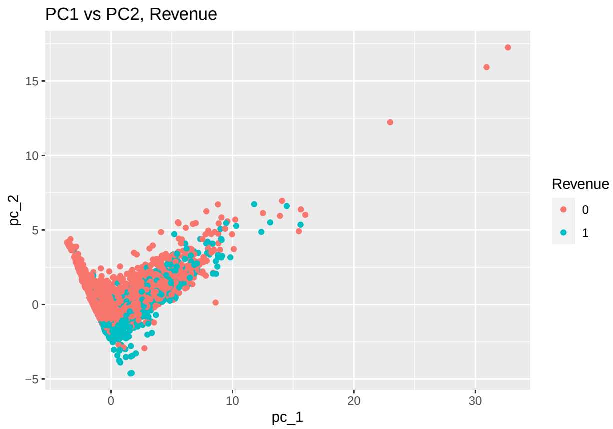
```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     0.07407934              0.02078606    0.01802058             0.01114437
## 2     0.10121780              0.02774738    0.02490882             0.01667445
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1     0.03685925              0.01527683  0.13180631 0.2417913 0.01397354
## 2     0.05581578              0.02318237  0.08331108 0.1803221 0.01933666
##   SpecialDay OperatingSystems  Browser   Region TrafficType  Weekend
## 1  0.1077536         2.165173 2.434201 3.147816    5.169156 1.227628
## 2  0.0000000         2.069421 2.254858 3.146765    2.611583 1.239200
##   Month_numeric VisitorType_Numeric
## 1      3.572628            1.211268
## 2      8.803622            1.375024
```

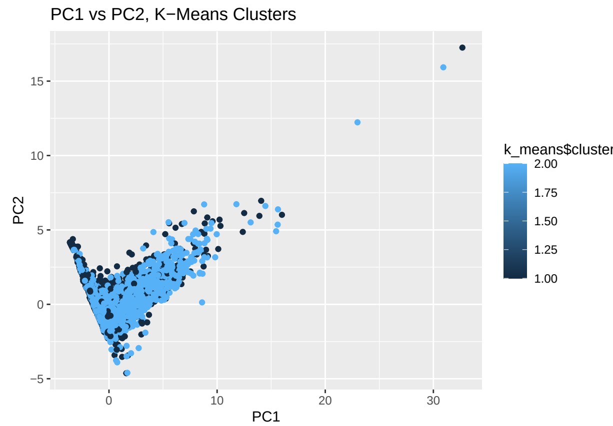
```r
#sum of squares
k_means$betweenss / k_means$totss
```

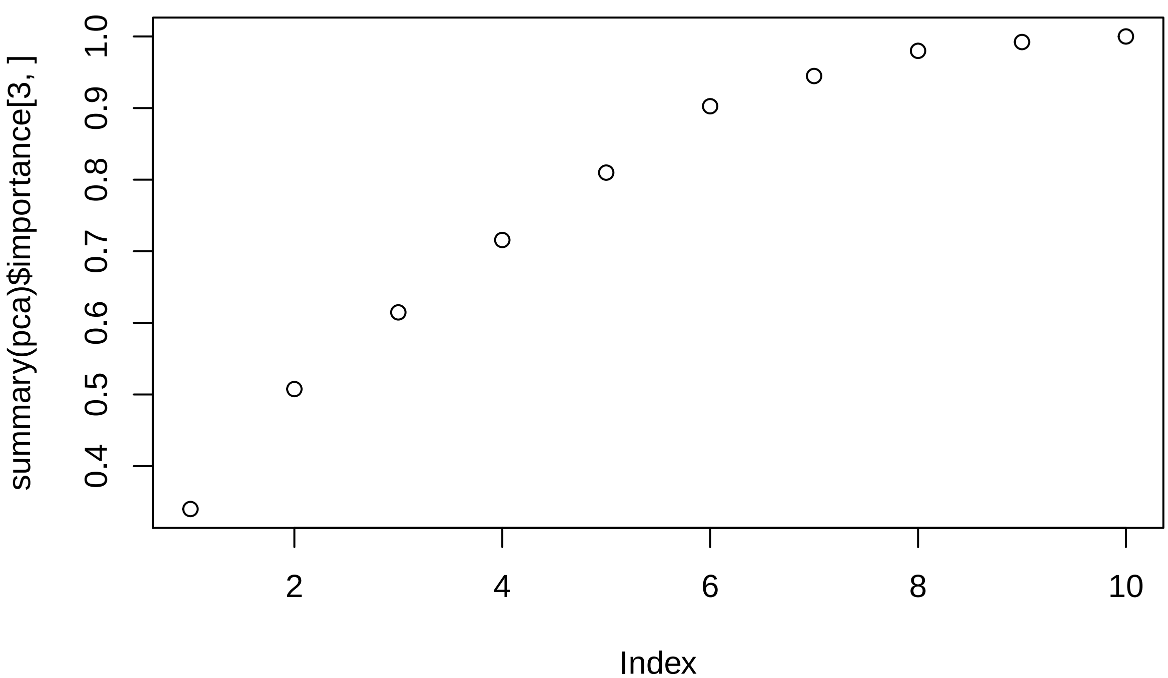
```
## [1] 0.2274077
```

A Cluster is a vector of integers 1:k indicating the cluster to which each point is allocated. Centers is a matrix of cluster centres. totss is the total sum of squares. withinss is a vector of within-cluster sum of squares, one component per cluster. tot.withinss is a total within cluster sum of squares. betweenss is between cluster sum of squares. The size represents the number of points in each cluster. K-means is a least-squares optimization problem. Principal Component Analysis(PCA) finds the least-squares cluster membership vector. Here, we use PCA to verify the clusters formed. PCA is used for dimensionality reduction, when the feature space contains too many irrelevant or redundant features. The aim is to find the intrinsic dimensionality of the data.

```
## Importance of components:
##                          PC1    PC2    PC3    PC4     PC5     PC6    PC7
## Standard deviation     1.844 1.2943 1.0350 1.0054 0.97009 0.96287 0.6496
## Proportion of Variance 0.340 0.1675 0.1071 0.1011 0.09411 0.09271 0.0422
## Cumulative Proportion  0.340 0.5076 0.6147 0.7158 0.80987 0.90258 0.9448
##                            PC8     PC9    PC10
## Standard deviation     0.59301 0.35055 0.27858
## Proportion of Variance 0.03517 0.01229 0.00776
## Cumulative Proportion  0.97995 0.99224 1.00000
```








```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
## 0.34004 0.50756 0.61469 0.71576 0.80987 0.90258 0.94478 0.97995 0.99224 1.00000
```

A cross-tabulation of Revenue type and cluster membership is given by

```
#confusion matrix
confusion_matrix <- table(k_means$cluster, scaling_data$Revenue)
confusion_matrix
```

```
##
##         0    1
```

```
##    1 6189  840
##    2 4233 1068
```

The confusion matrix is one of the most intuitive metric used for finding the correctness and accuracy of the model.The ideal scenario would be that the model should give 0 False Positives and 0 False Negatives. But that's not the case in real life as any model will not be 100% accurate most of the times. We know that there will be some error associated with every model that we use for predicting the true class of the target variable. The predictive power of the model is determined by three measures precision, recall and F1 score. Precision is a good measure to determine, when the costs of False Positive is high. Recall calculates how many of the actual Positives our model capture through labeling it as true Positive. F1 score is the best measure which balances between precision and recall and when there is a uneven class distribution.

```
#predictive power of the model
precision_kmeans<- confusion_matrix [1,1]/(sum(confusion_matrix [1,]))
precision_kmeans
```

```
## [1] 0.8804951
```

```
recall_kmeans<- confusion_matrix [1,1]/(sum(confusion_matrix [,1]))
recall_kmeans
```

```
## [1] 0.59384
```

```
#F1 score
F1<- 2*precision_kmeans*recall_kmeans/(precision_kmeans+recall_kmeans)
F1
```

```
## [1] 0.7093003
```

The model depicts high error rates and low F1 score. We can try with centers = 4.

```
str(k_means_4)
```

```
## List of 9
##  $ cluster     : int [1:12330] 1 1 1 1 1 1 1 1 1 1 ...
##  $ centers     : num [1:4, 1:17] 0.0736 0.0944 0.0735 0.1037 0.0203 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "1" "2" "3" "4"
##   .. ..$ : chr [1:17] "Administrative" "Administrative_Duration" "Informational" "Informational_Dura
##  $ totss       : num 451708
##  $ withinss    : num [1:4] 54521 22040 61827 36309
##  $ tot.withinss: num 174697
##  $ betweenss   : num 277011
##  $ size        : int [1:4] 5064 1477 1823 3966
##  $ iter        : int 3
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
#size of cluster
k_means_4$size
```

```
## [1] 5064 1477 1823 3966
```

```r
#Means
k_means_4$centers
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     0.07361331              0.02025080    0.01825796             0.01149125
## 2     0.09438552              0.02503368    0.02079102             0.01187737
## 3     0.07346458              0.02158286    0.01714207             0.01067755
## 4     0.10366822              0.02882595    0.02629644             0.01803463
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1     0.03531954              0.01457182  0.12275897 0.2339038 0.01364625
## 2     0.04573004              0.01928027  0.08382403 0.1825187 0.01760081
## 3     0.04249950              0.01790010  0.16575074 0.2717727 0.01486683
## 4     0.05826654              0.02404691  0.08080572 0.1779948 0.01979840
##   SpecialDay OperatingSystems  Browser   Region TrafficType  Weekend
## 1 0.11749605         2.063389 2.357622 2.761651    2.719984 1.225513
## 2 0.02288422         2.092756 2.373053 7.377793    2.530129 1.234936
## 3 0.07054306         2.459133 2.629731 3.319803   12.617115 1.235875
## 4 0.00000000         2.059002 2.225164 1.985124    2.437216 1.239284
##   Month_numeric VisitorType_Numeric
## 1      2.779028            1.212875
## 2      7.474611            1.406906
## 3      6.397148            1.177180
## 4      8.826273            1.370903
```

```r
#sum of squares
k_means_4$betweenss / k_means_4$totss
```

```
## [1] 0.6132531
```

```r
#confusion matrix
confusion_matrix_4 <- table(k_means_4$cluster, scaling_data$Revenue)
confusion_matrix_4
```

```
##
##        0    1
##   1 4503  561
##   2 1224  253
##   3 1565  258
##   4 3130  836
```

```r
#predictive power of the model
presicion_kmeans_4<- confusion_matrix_4 [1,1]/(sum(confusion_matrix_4[ 1,]))
presicion_kmeans_4
```

```
## [1] 0.889218
```

```r
recall_kmeans_4<- confusion_matrix_4[1,1]/(sum(confusion_matrix_4[,1]))
recall_kmeans_4
```

```
## [1] 0.4320668
```

```r
#F1 score
F1_4<- 2*presicion_kmeans_4*recall_kmeans_4/(presicion_kmeans_4+recall_kmeans_4)
F1_4
```

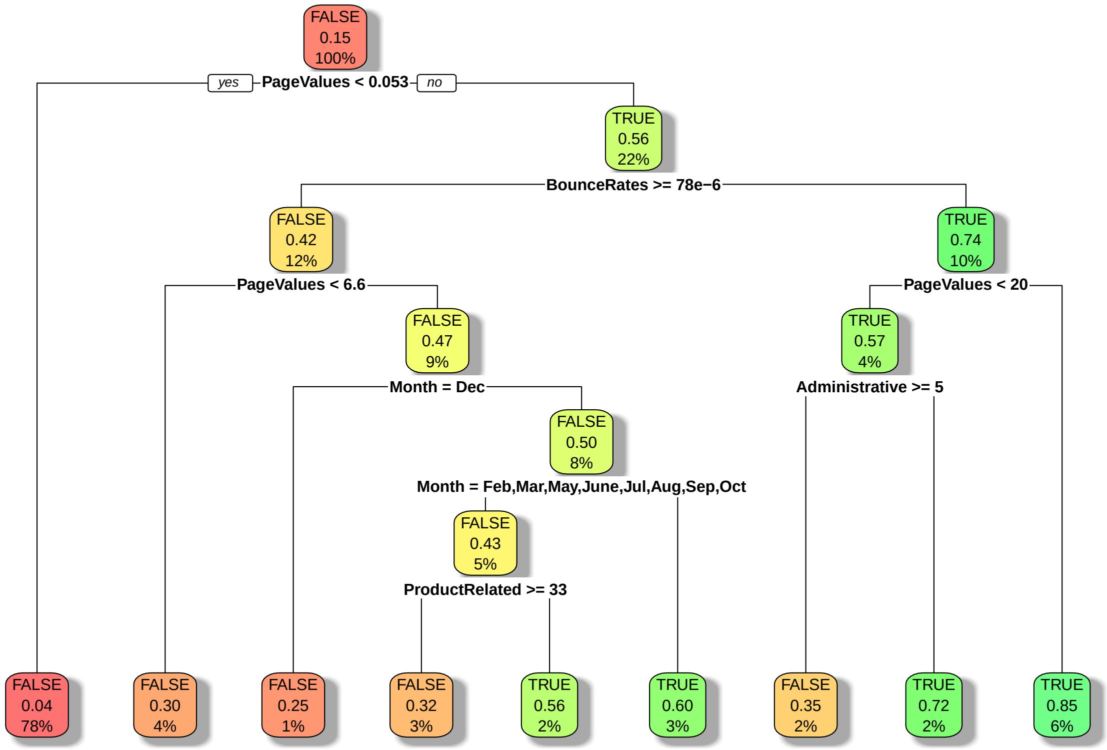
```
## [1] 0.5815575
```

F1 score reveals very little change. Clustering techniques did not observe any significant performance improvement. Due to class imbalance problem we were not able to perform in clustering models. Hence we may need more data to perform better. Next, we will try decision tree model. Decision tree is a widely used classifier. The first use of a tree-based decision system was used in artificial intelligence in 1960. Decision tree analyzes and extracts valuable rules as well as relationships from large data source. Since the decisions are made at multiple levels, these supervised classifiers are more efficient than single stage classifiers. It uses tree structure to make decisions. Tree structure consists of root node, child nodes and leaf nodes; each node makes decision based on its attribute value of data. One of the most commonly used decision tree is binary tree uses tree growing approach for classification. In binary trees, a case traversing to the left child is true while a case traversing to the right is false. When more features are introduced, the problem of classification becomes much more complex. The difficulty in utilizing decision trees lies in their construction. Here is the data we are going to use:

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Ord.factor w/ 10 levels "Feb"<"Mar"<"May"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ OperatingSystems       : Ord.factor w/ 8 levels "6"<"3"<"7"<"1"<..: 4 6 7 2 2 6 6 4 6 6 ...
##  $ Browser                : Ord.factor w/ 13 levels "9"<"3"<"6"<"7"<..: 5 6 5 6 2 6 9 6 6 9 ...
##  $ Region                 : Ord.factor w/ 9 levels "8"<"6"<"3"<"4"<..: 6 6 9 8 6 6 3 6 8 6 ...
##  $ TrafficType            : Ord.factor w/ 20 levels "12"<"15"<"17"<..: 9 16 7 11 11 7 7 15 7 16 ...
##  $ VisitorType            : Ord.factor w/ 3 levels "Returning_Visitor"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weekend_01             : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue_01             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```r
#Accuracy
mean(predict_dt==test_data$Revenue_01) #Accuracy
```

```
## [1] 0.8999459
```




```
##                            fit_dt.variable.importance
## PageValues                                883.6872939
## BounceRates                               103.6037944
## ProductRelated                             71.3877504
## Administrative                             67.1077807
## ProductRelated_Duration                    53.5209703
## ExitRates                                  38.0443428
## VisitorType                                26.9744430
## Informational_Duration                     21.7469218
## Month                                      17.6908407
## Administrative_Duration                    14.1352182
## Informational                              12.4279250
## TrafficType                                 0.2017293
```

From the above decision tree, it is evident that the most significant attribute contributing towards the most information output. The variable importance table describes all the revenue drivers. The F1 Score is considerable increase as compared to previous models. PageValue suggests that customers look at different variety of products. So optimization of website pages is very important. Personalized tracking of customers, reducing the exit rate, engaging the new visitors, Weekend promotional activities, Festive season discounts and offers, User friendly website, working on marketing strategy, promoting via social media, a good recommendation system for suggesting variety of products can improve the revenue drastically.

RESULT

The evaluation metrics are precision, recall, F1 score. The decision tree gave a very precise model (0.92) that also has good recall (0.96) and high F1 score value of 0.94. The final prediction accuracy is 0.89. Thus the decision tree model is a powerful predictive tool when compared to clustering technique because of the limited data.

```r
#Predictive power of the decision tree model
confusion_matrix_dt<- table(predict_dt,test_data$Revenue_01)
confusion_matrix_dt
```

```
##
## predict_dt FALSE TRUE
##      FALSE  3007  251
##      TRUE    119  321
```

```r
#Precision
presicion_dt<- confusion_matrix_dt[1,1]/(sum(confusion_matrix_dt[1,]))
presicion_dt  #Precision
```

```
## [1] 0.9229589
```

```r
#Recall
recall_dt<- confusion_matrix_dt[1,1]/(sum(confusion_matrix_dt[,1]))
recall_dt #Recall
```

```
## [1] 0.9619322
```

```r
#F1 score
F1_dt<- 2*presicion_dt*recall_dt/(presicion_dt+recall_dt)
F1_dt  #F1 score
```

```
## [1] 0.9420426
```

CONCLUSION

In this project we predict, based on an extensive set of predictors from different categories, whether a potential customer will engage in online-purchasing behaviour. Though our dataset is limited in size, we are able to highlight the list of suggestions via decision tree model which may improve e-retailers target. We can also examine whether the results only hold for small e-commerce companies or can be generalized to all shops should be tested in additional studies. The prediction accuracy, especially in the recognition of a few categories, needs to be improved. In the future, in-depth research can be made on the prediction of purchases of multiple categories of products, making real-time predictions and personalization of users browsing preferences.

REFERENCES

1) CITATION REQUEST FOR DATA: Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)

2) Introduction to Data Science - Data Analysis and Prediction Algorithms with R by Prof. Rafael A. Irizarry (2019-03-27)

3) Xiaotong Dou, Online Purchase Behavior Prediction and Analysis Using Ensemble Learning , Institute of Electrical and Electronics Engineers(IEEE), 5th International Conference on Cloud Computing and Big Data Analytics,2020

4) Dirk Van den Poel,Wouter Buckinx, Predicting online-purchasing behaviour, European Journal of Operational Research(ELSEVIER), Interfaces with Other Disciplines, 2004