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
```

```
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
```

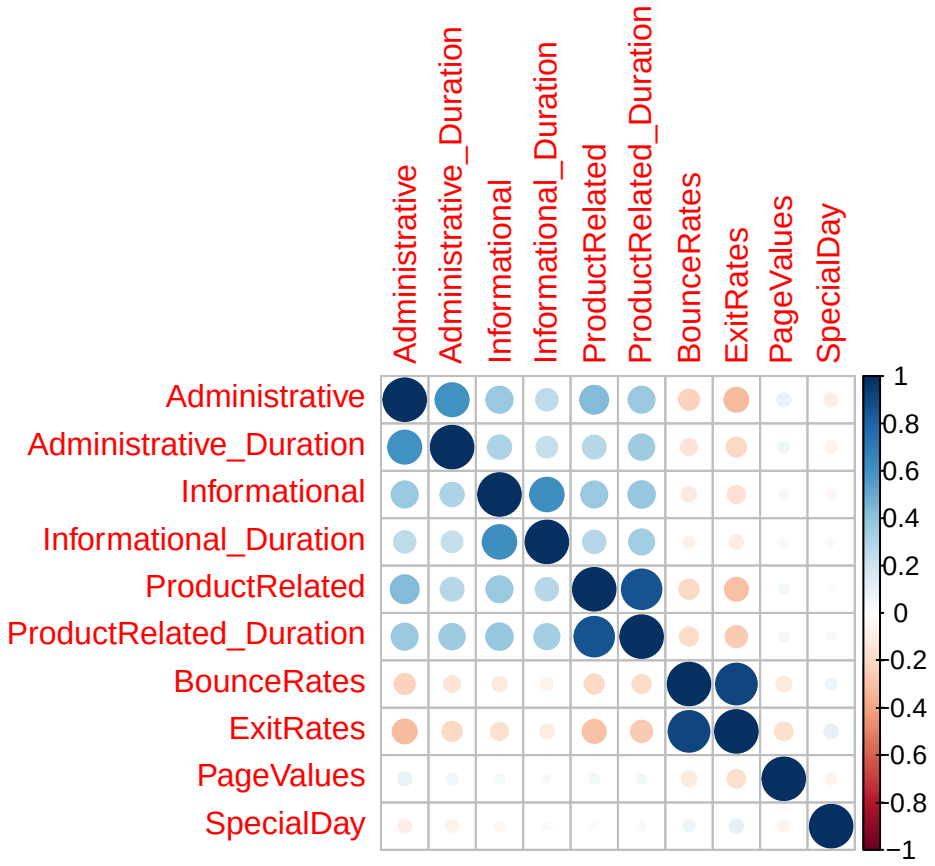
The summary statistics of Product_Related is

```
##    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

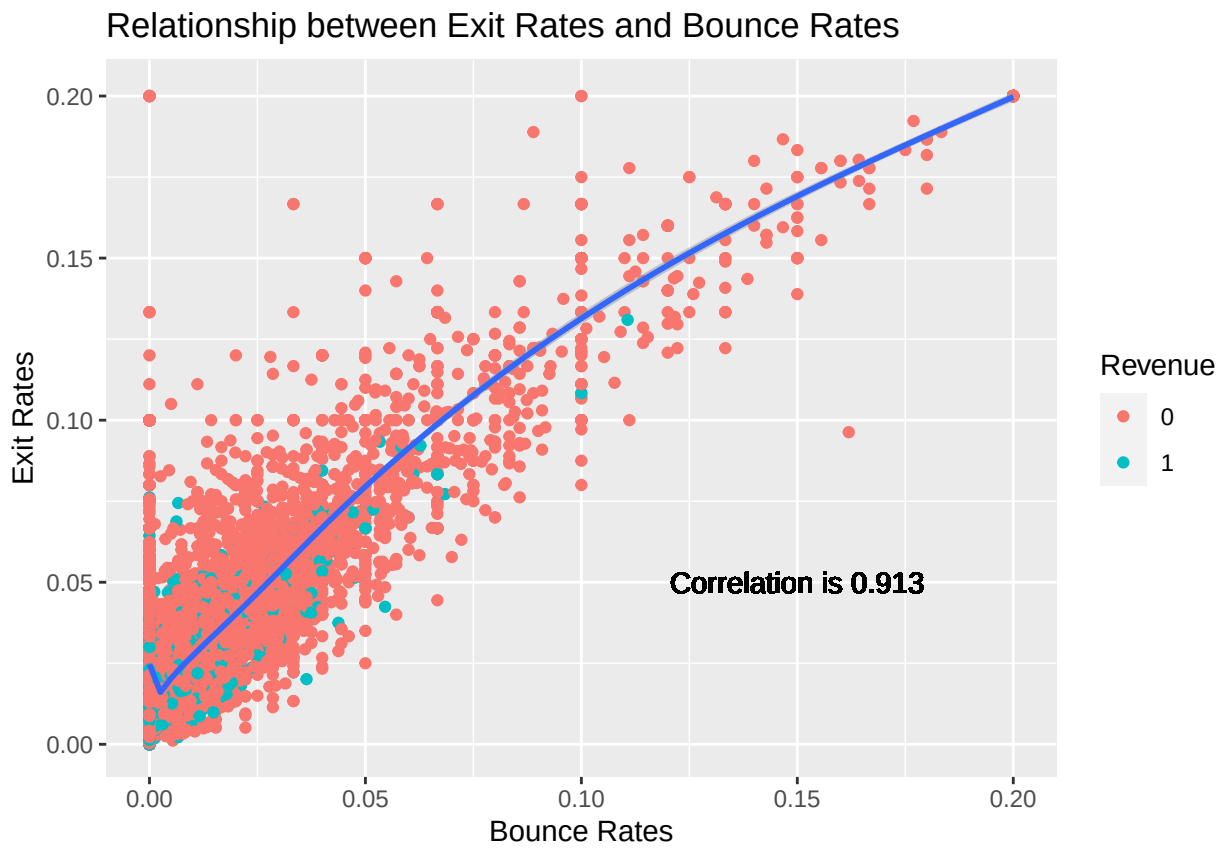
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   184.1   598.9  1194.8  1464.2 63973.5
```

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

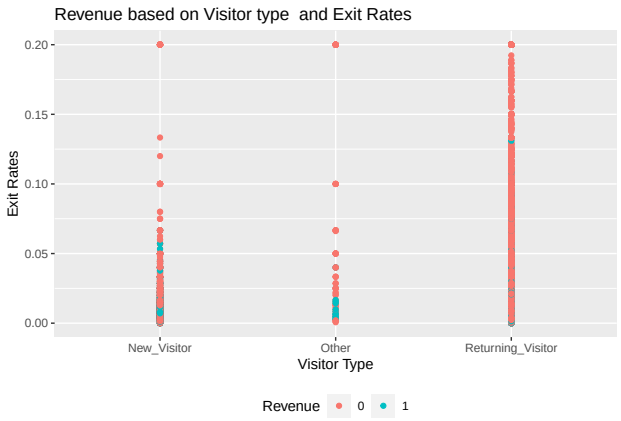


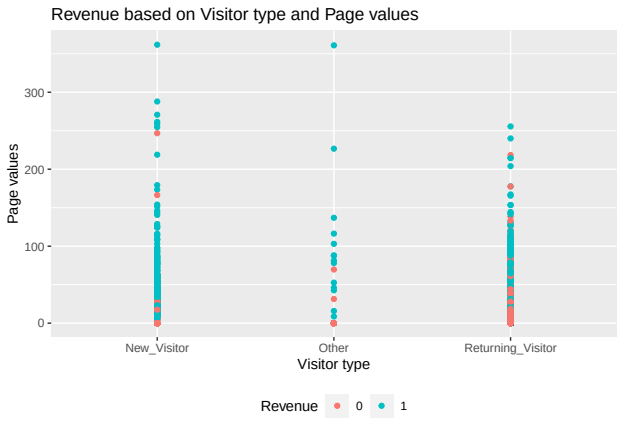
Let us plot the relationship between Bounce Rates and Exit Rates. It is evident from the plot, the shoppers who exit early are some of our potential customers. It is wise show some attractive pop ups like discount or huge offer when a customer attempt to leave the site.

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

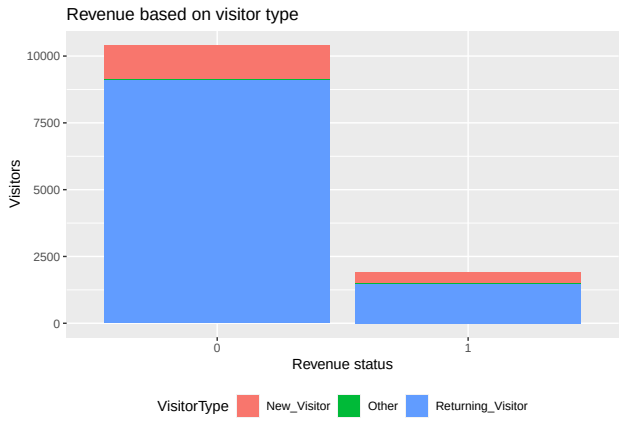


When we explore the relationship between visitor type-Exit Rate and visitor type-page values with respective to Revenue, the new visitor contributes more revenue than the returning visitor. Offering the reference coupons and giving discounts on it can bring new customers.

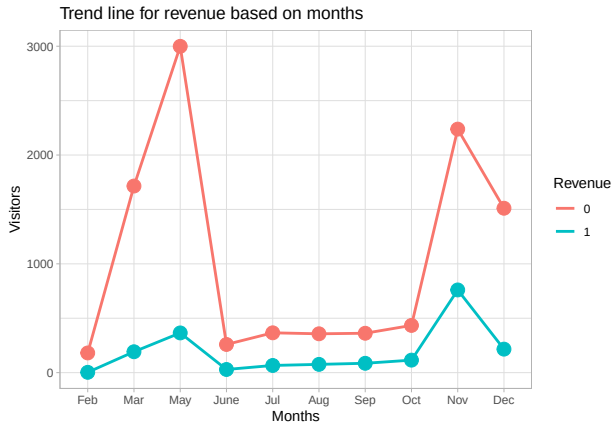


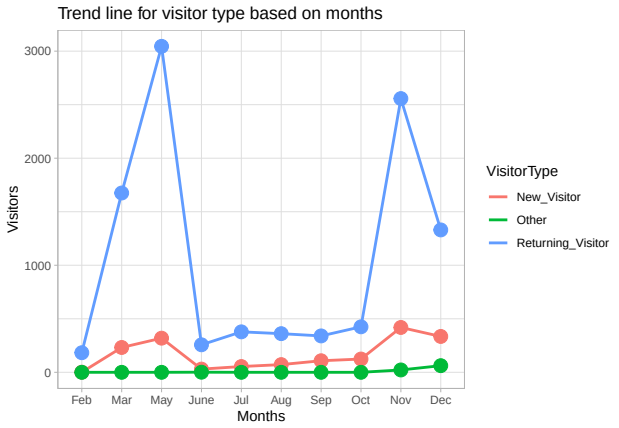


The conversion rate of potential customers is very important. Concentrating on new customers will significantly improve the sales and revenue growth. From the below plot, the purchase made during the weekday is higher than the weekends. Introducing weekends based promotional events may help the shoppers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



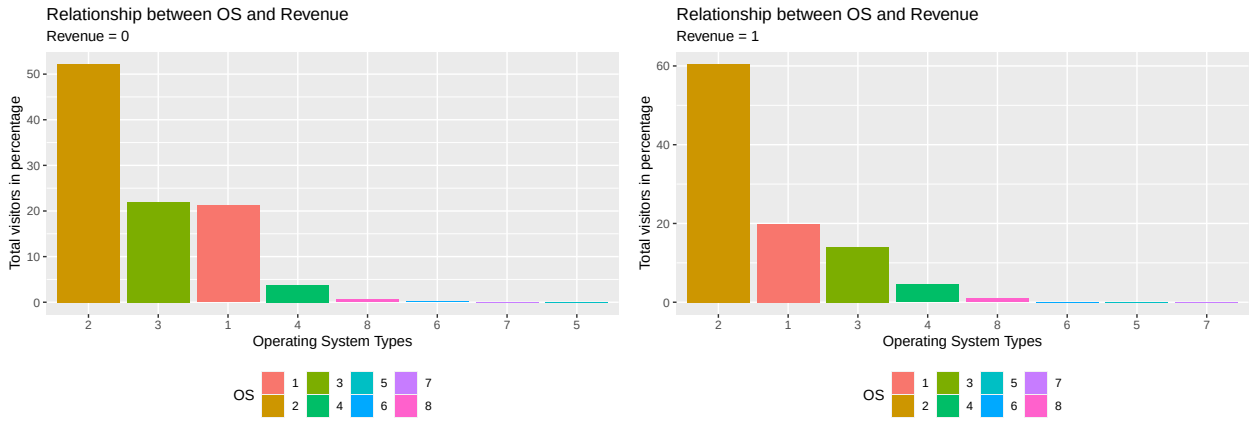
The below plot explain the seasonality revenue improvement. There seems to be many customers buy products during March to May and October to November. The plot also suggests that lot of customer are viewing the item but final transactions are made after adding into the cart. There may be hidden charges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





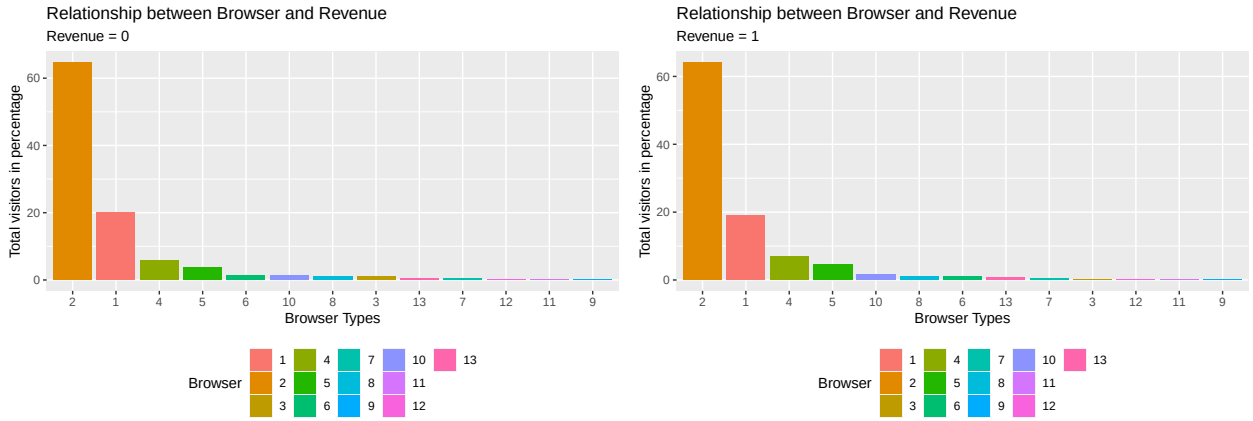
The operating systems of the user may also be considered as significant characteristics of predicting the shoppers. Most of our customer uses '2' OS type. Other OS are used by less customers. This could also mean many customer are not preferring to use the site in other sources.

```
## '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 ...
```



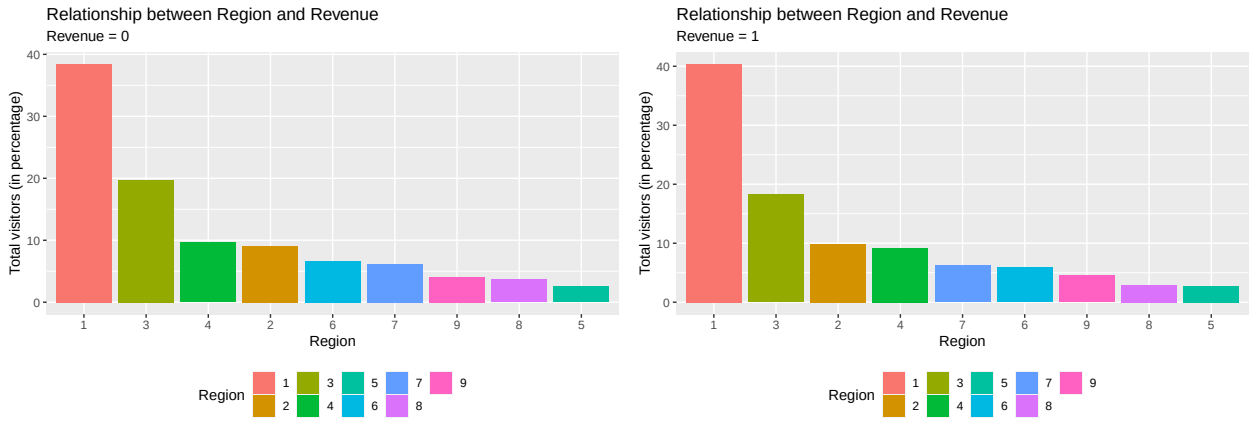
The relationship between Browser and Revenue states that the type '2' remains at the top. This may also suggest the website is not user friendly with other type of browsers. Web designers can concentrate on this for better improvement.

```
## '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 ...
```



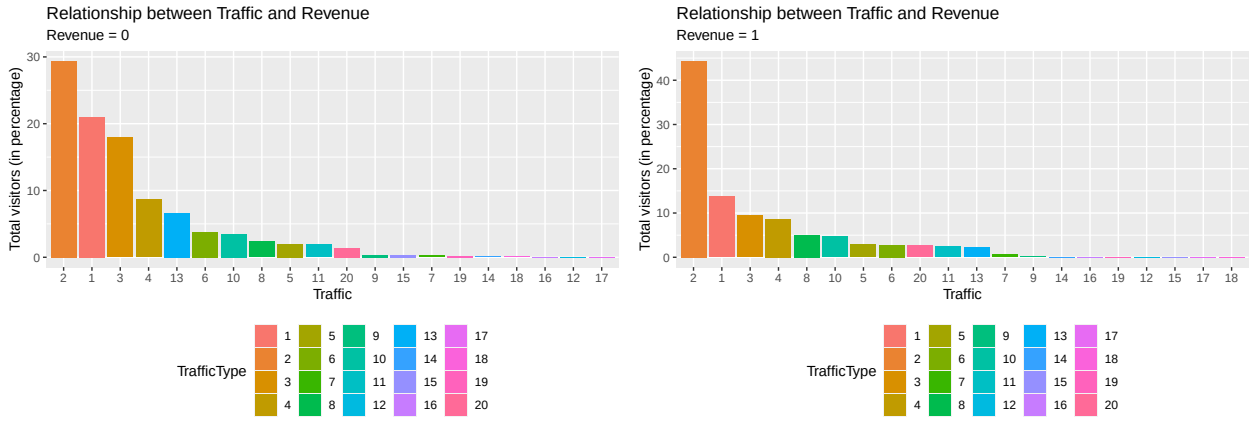
The relationship between Region and Revenue states that the most of our customers are from '1' and '3'. The marketing reach strategy can be helpful in these regions.

```
## '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 ...
```