ONLINE SHOPPING INTENTION

SUBMITTED BY:
Group No.1

Batch:
August2019-Hyderabad

By:
- Arul Vishwakarma

UNDER THE GUIDANCE OF:
Mr. Srikar Muppidi

PGPDSE

By

Great Lakes Institute of Management

CERTIFICATE

This is to certify that the Project Report entitled **ONLINE SHOPPING INTENTION**which is submitted by **Arul, Sai Krishna, Praveen, Abha** and**Suneetha**in partial fulfillment of the requirement for the award of Post Graduate Program in Data Science and Engineering (PGPDSE), is a record of the candidates own work carried out by him under my supervision.The matter embodied in this is original and has not been submitted for the award of any other degree.

Mr. SRIKAR MUPPIDI

Date:07 /1/2020

Place: Hyderabad

DECLARATION

I declare that the project entitled **ONLINE SHOPPING INTENTION**is a project work carried out by us under the supervision and guidance of **Mr. SRIKAR MUPPIDI** for the award of project degree PGPDSE, and this has not been previously submitted for the award of any Degree, Diploma or other similar title of any other University/ Institute.

Date:07/1/2020
Place: Hyderabad

Group No.1:

Arul
Sai Krishna
Praveen
Abha Jain
Suneetha

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor **SRIKAR MUPPIDI** for providing his invaluable guidance, comments and suggestions throughout the course of this project. We value the assistance of Great Learning, Hyderabad campus. Learning from their knowledge helped me to become passionate about my research topic.

We will be failing in our duty if each one of us don't express our gratitude for other team members for the valuable contributions during course of this project.

Group No.1:

Arul
Sai Krishna
Praveen
Abha Jain
Suneetha

ABSTRACT

In this paper, we propose a real-time online shopper behavior analysis system consisting of two modules which simultaneously predicts the visitor's shopping intent and Web site abandonment likelihood. In the first module, we predict the purchasing intention of the visitor using aggregated pageview data kept track during the visit along with some session and user information. The extracted features are fed to random forest (RF), support vector machines (SVMs) classifiers as input. We use oversampling and feature selection preprocessing steps to improve the performance and scalability of the classifiers. The results show that MLP that is calculated using resilient backpropagation algorithm with weight backtracking produces significantly higher accuracy and F1 Score than RF and SVM. Another finding is that although clickstream data obtained from the navigation path followed during the online visit convey important information about the purchasing intention of the visitor, combining them with session information-based features that possess unique information about the purchasing interest improves the success rate of the system. In the second module, using only sequential clickstream data, we train a long short-term memory-based recurrent neural network that generates a sigmoid output showing the probability estimate of visitor's intention to leave the site without finalizing the transaction in a prediction horizon. The modules are used together to determine the visitors which have purchasing intention but are likely to leave the site in the prediction horizon and take actions accordingly to improve the Web site abandonment and purchase conversion rates. Our findings support the feasibility of accurate and scalable purchasing intention prediction for virtual shopping environment using clickstream and session information data

TABLE OF CONTENT

CHAPTER-1

INTRODUCTION

The increasing popularity of online shopping has led to the emergence of new economic activities. To succeed in the highly competitive e-commerce environment, it is vital to understand consumer intention. Understanding what motivates consumer intention is critical because such intention is key to survival in this fast-paced and hypercompetitive environment. Where prior research has attempted at most a limited adaptation of the information system success model, we propose a comprehensive, empirical model that separates the ‘use’ construct into ‘intention to use’ and ‘actual use’. This makes it possible to test the importance of user intentions in determining their online shopping behaviour. Our results suggest that the consumer's intention to use is quite important, and accurately predicts the usage behaviour of consumers. In contrast, consumer satisfaction has a significant impact on intention to use but no direct causal relation with actual use.

Problem statement:

This dataset represents sessions of users on a website.The features encode the online activity attributes of the user captured in his/her each session.The task is to predict whether a user in that session is going to buy something from the website (generate Revenue) or not.The features that would look for in prospective customer to ensure a buying. The most important features defining the buy/not buy is the intent of customer.

Data-set: The dataset consists of 18 feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

Shape:12,330 rows and 18 columns

CHAPTER-2

LITERATURE

Feature Name	Type	Description
Administrative	Numeric	Number of pages by the visitor about account management
Administrative Duration	Numeric	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Numeric	Number of pages visited by the visitor about Web site, communication and address information of the shopping site
Informational Duration	Numeric	Total amount of time (in seconds) spent by the visitor on informational pages
Product Related	Numeric	Number of pages visited by visitor about product related pages
Product Related Duration	Numeric	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce Rate	Numeric	Average bounce rate value of the pages visited by the visitor
Exit Rate	Numeric	Average exit rate value of the pages visited by the visitor
Page Value	Numeric	Average page value of the pages

		visited by the visitor
Special Day	Numeric	Closeness of the site visiting time to a special day
Month	Nominal	Month of the year
OperatingSystems	Numeric	Operating system used
Browser	Numeric	Browser used
Region	Numeric	Region of the user
TrafficType	Numeric	Traffic Type
VisitorType	Nominal	Types of Visitor
Weekend	Boolean	Weekend or not
Revenue	Boolean	Revenue will be generated or not

CHAPTER-3
DATA CLEANING

The following changes have been done for better analysis, visualization and model building. The changes done for the required columns are as below:

Missing Values:

We can start analysis by looking at the percentage of missing values in each column. Missing values are fine when we do Exploratory Data Analysis, but they will have to be filled in for machine learning methods.

```
Administrative             0.0
Administrative_Duration    0.0
Informational              0.0
Informational_Duration     0.0
ProductRelated             0.0
ProductRelated_Duration    0.0
BounceRates                0.0
ExitRates                  0.0
PageValues                 0.0
SpecialDay                 0.0
Month                      0.0
OperatingSystems           0.0
Browser                    0.0
Region                     0.0
TrafficType                0.0
VisitorType                0.0
Weekend                    0.0
Revenue                    0.0
dtype: float64
```

This says that in this dataset there are no missing values.

Convert Object Type to Numerical Types:

We convert the columns with numbers into numeric data types by replacing the strings which can be interpreted as floats. Then we will convert the columns that contain numeric values into numeric data types.

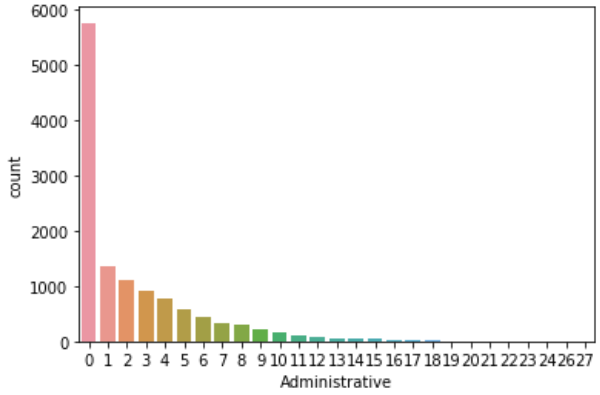
CHAPTER-4

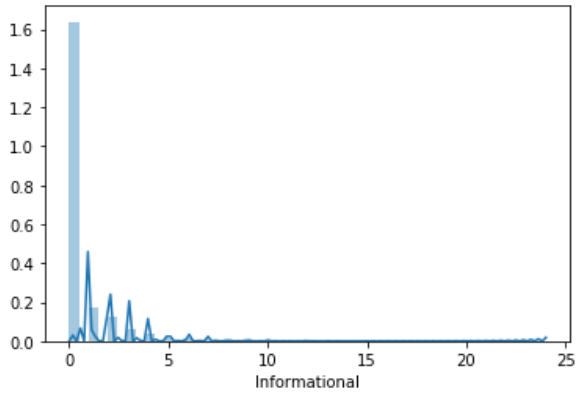
EXPLORATORY DATA ANALYSIS

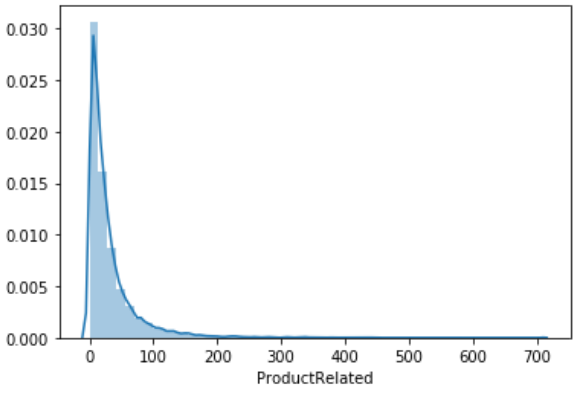
Exploratory Data Analysis (EDA) is an open-ended process where we make plots and calculate statistics in order to explore our data. The purpose is to find anomalies, patterns, trends, or relationships. These may be interesting by themselves (for example finding a correlation between two variables) or they can be used to inform modeling decisions such as which features to use. In short, the goal of EDA is to determine what our data can tell us. EDA generally starts out with a high-level overview, and then narrows into specific parts of the dataset once as we find interesting areas to examine.

UNIVARIATE ANALYSIS:

TYPES OF PAGES:

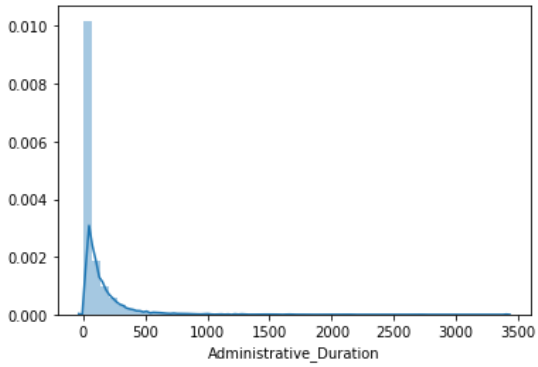


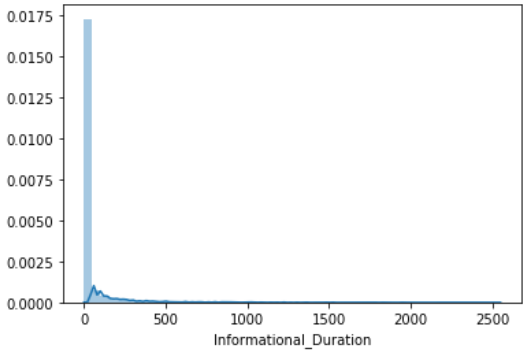


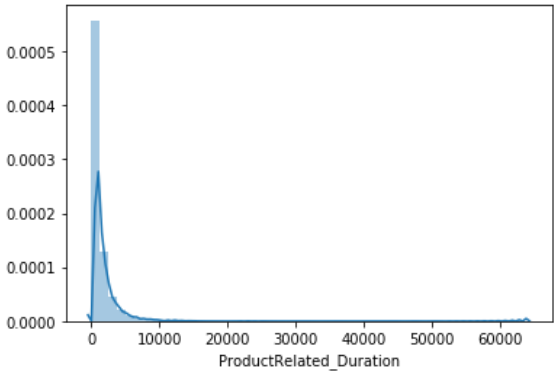


Different types of pages when a visitor visits a site.These are the different type of pages present in any website .As we can see that administrative page of type 1 is accessed more.From the informative page we can see there are more visitors for type-0.As for the product related more visits are done between 0-100.

DURATION OF THE TYPES OF PAGES:



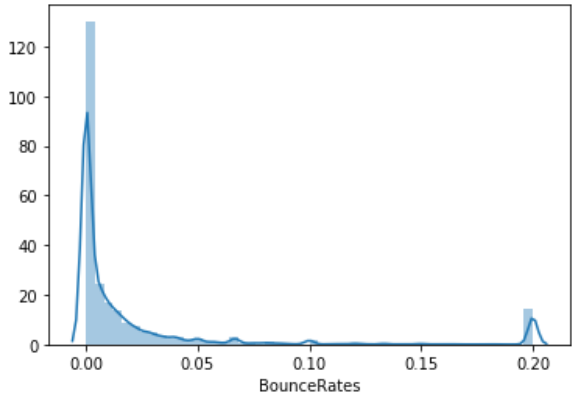




When a visitor visits a site and spending some time in a particular page type

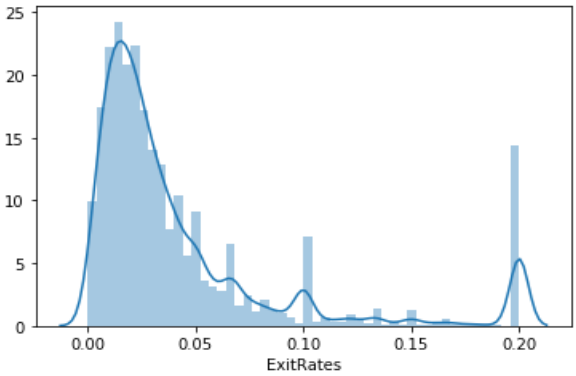
That page contains a duration page and can predict that person is buying are not.

BOUNCE RATES:



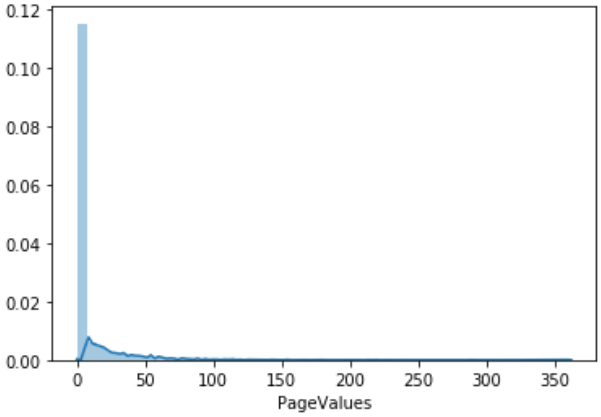
Bounce rate is the **percentage** of single page visits (or web sessions). It is the **percentage** of visits in which a person leaves your website from the landing page without browsing any further.

EXIT RATES:



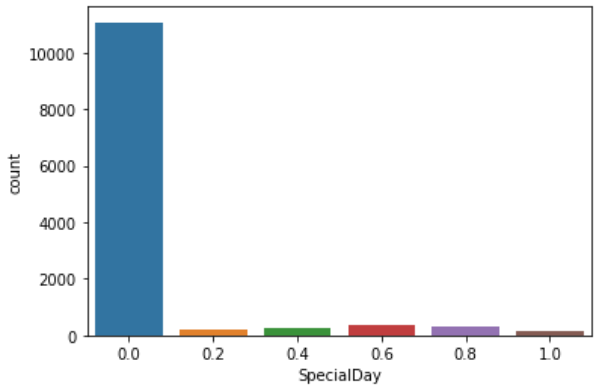
The **exit rate** is the **percentage** that were the last in the session. For all sessions that start with the page, bounce **rate** is the **percentage** that were the only one of the session. The bounce **rate** calculation for a page is based only on visits that start with that page.

PAGE VALUE:



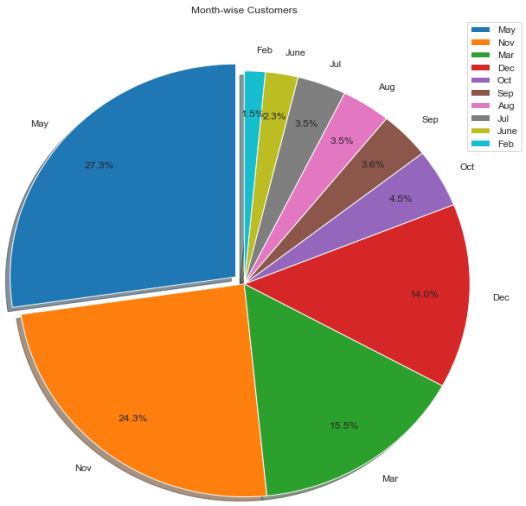
Page Value is the average **value** for a **page** that a user visited before landing on the goal **page** or completing an Ecommerce transaction (or both).

SPECIAL DAY:



Special Day range shows the range from 0.0 to 1.0. Closeness of the site visiting time to a special day (holiday).

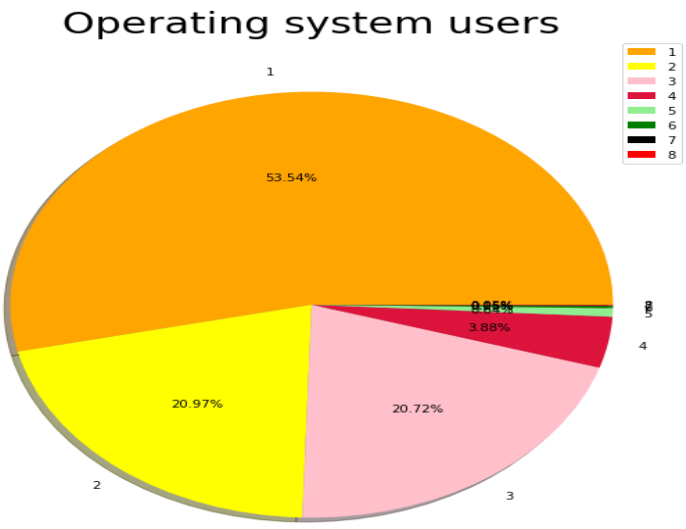
MONTH:



The above plot describes the no visitors the site in a particular month. 27% customers are visiting in May, 24% in Nov and 15% in March. So there is no specific month when customers visit the sites mostly. We can see that most of the customers visit the website on November and December itself, which can be during the holidays.

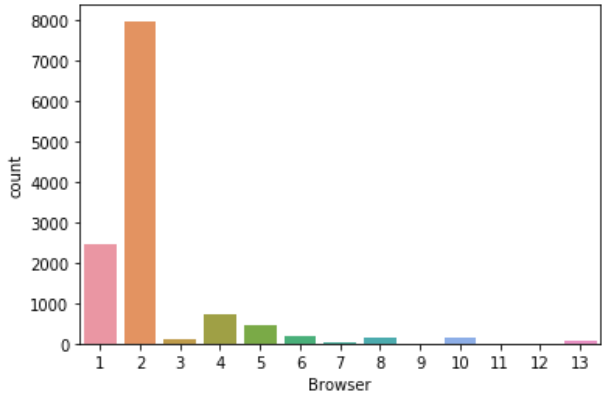
OPERATING SYSTEMS:

	Operating System
☐	1. Windows
☐	2. iOS
☐	3. Android
☐	4. Android
☐	5. Windows
☐	6. Windows
☐	7. Macintosh
☐	8. Android
☐	9. iOS



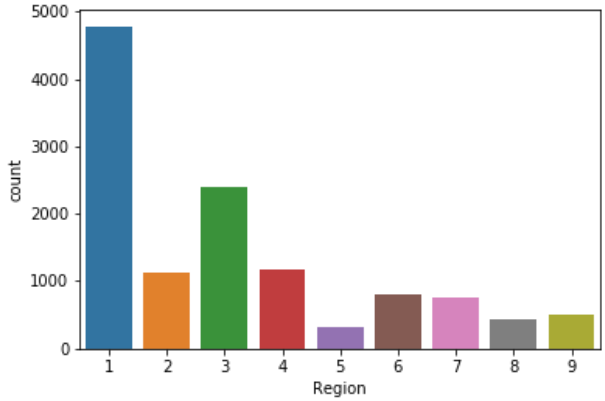
Most of the visitors visits the site through the same type of operating system.

BROWSER:



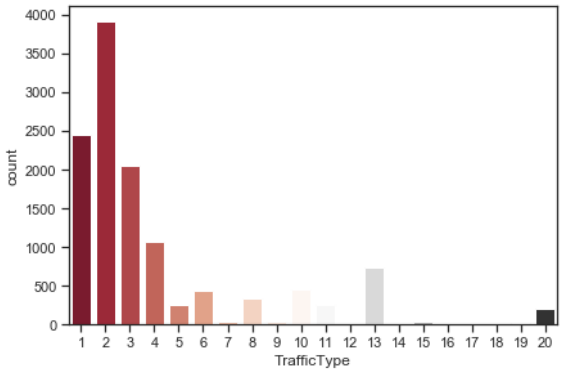
Most of the visitors visits the site using the same type of Operating system.

REGION:



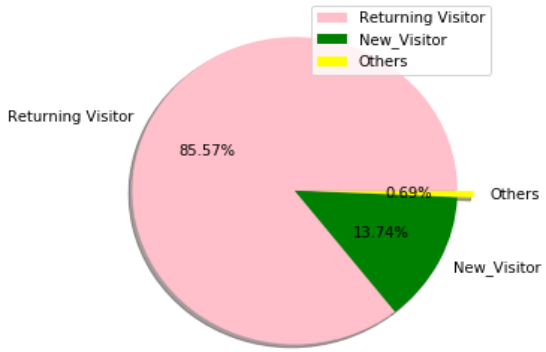
Most of the visitors visits the site who belongs to the same Region.

TRAFFICTYPE:



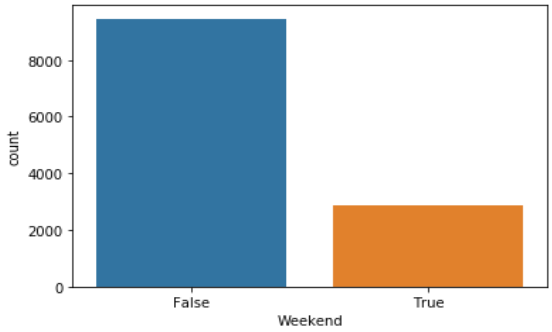
Most of the visitors visits the site through the same way(directly visiting the site or by some advertisements).

VISITERTYPE:



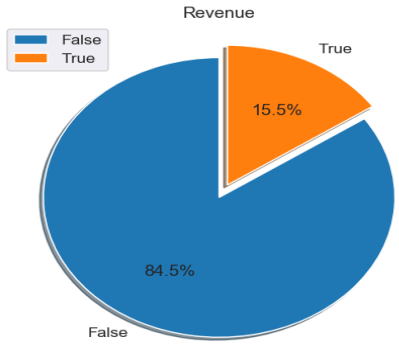
The plot describes the mostly returning visitor(regular visitors),compared to new visitor (who are new registers of the site) and other visitors (who are visiting privately or unregistered visitors)**.**

WEEKEND:



Probability of the customers whether buying in weekends or not on the above plot clearly describes the customers that they are buying are not True means buying they are buying False means not buying.

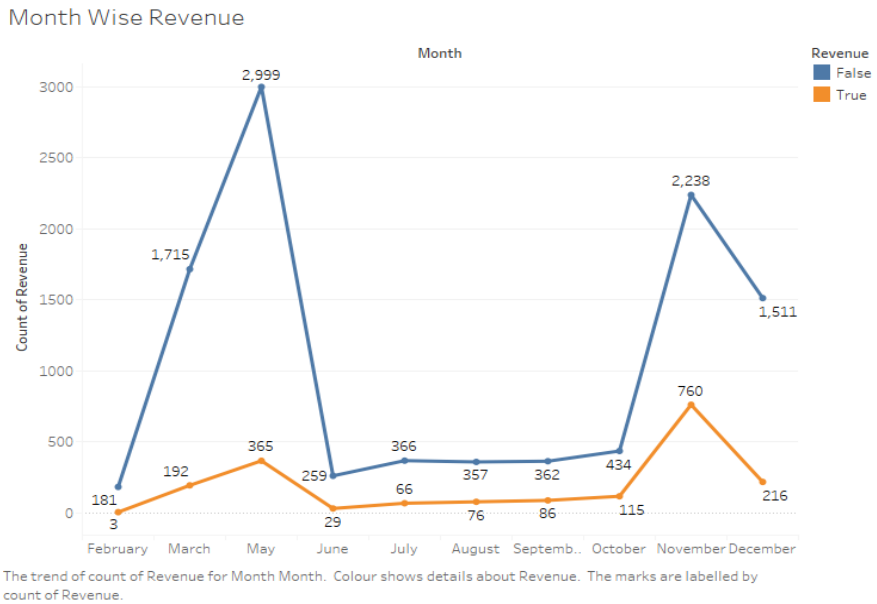
REVENUE:



Only 15% of total customers visiting the shopping sites are giving final revenue.

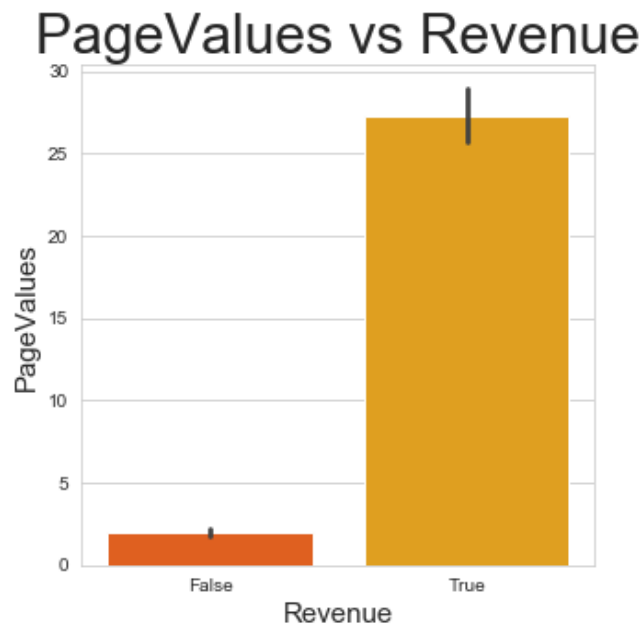
BIVARIATE ANALYSIS:

MONTH WITH REVENUE:



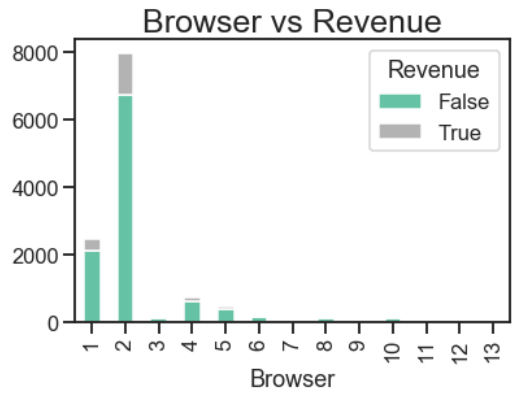
In the Month of November and May the Revenue is high.

PAGEVALUE WITH REVENUE:



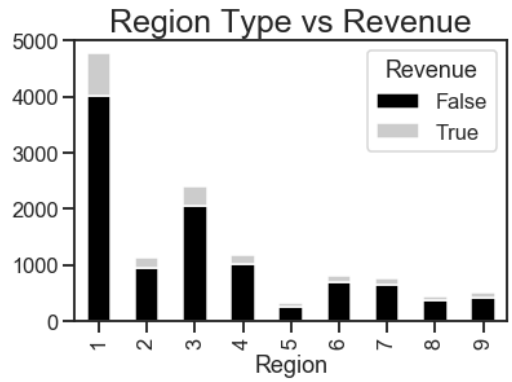
Most of the Revenue lies in Page Value.

BROWSER WITH REVENUE:



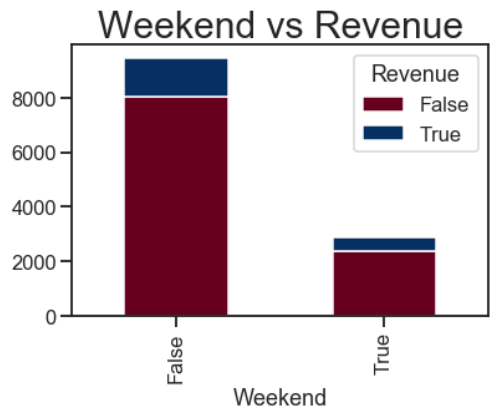
Most of the Revenue falls under the visitors who visits the site through the same type of browser.

REGION WITH REVENUE:



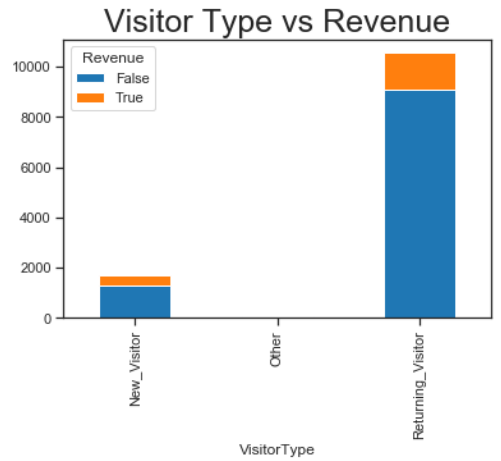
Most of the Revenue falls under the visitors who visits the site that belongs to the same region.

WEEKEND WITH REVENUE:



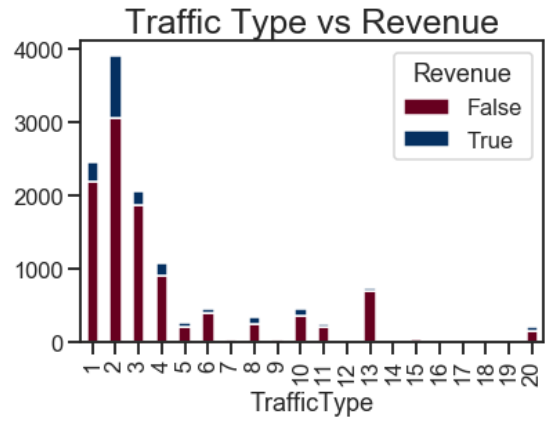
Revenue is more in weekdays.

VISITORTYPE WITH REVENUE:



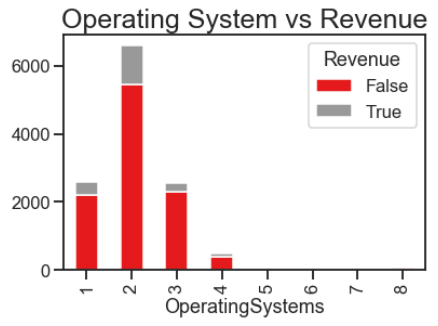
The Returning visitor are giving more revenue to the shopping site.

TRAFFIC TYPE WITH REVENUE:



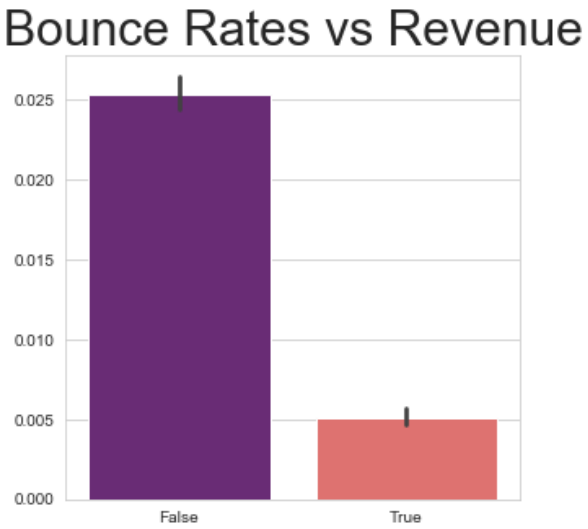
Most of the visitors visits the site through the above plot describing the visitors visiting the site through the same way(directly visiting the site or by some advertisements)

OPERATING SYSTEM WITH REVENUE:



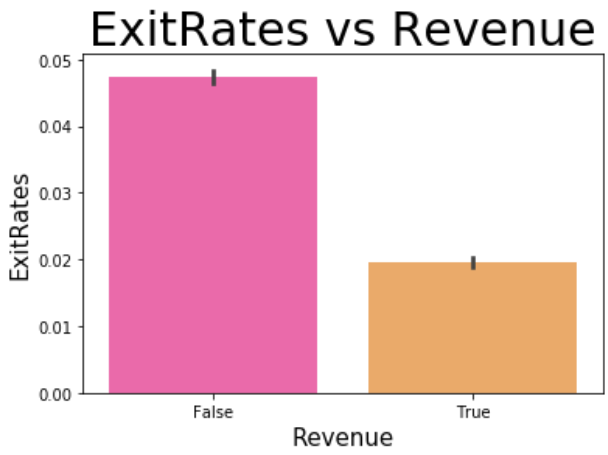
Most of the visitors visits the site through the same type of operating system gives more Revenue.

BOUNCE RATE WITH REVENUE:



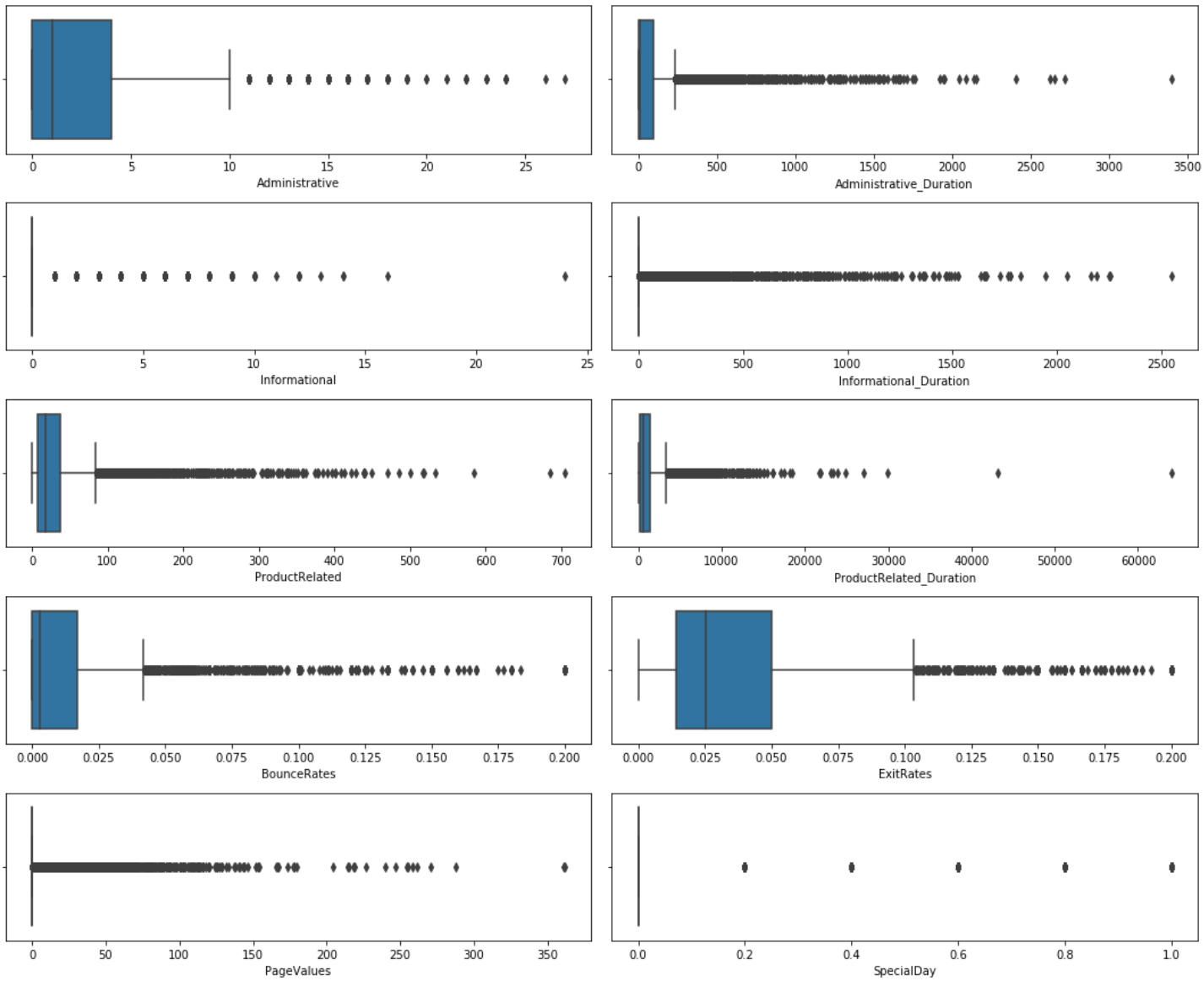
As we can see that bounce rate is less for the one who generated the revenue.

EXIT RATE WITH REVENUE:



For all page views to the page, the exit rate is the percentage that were the lastin the session.

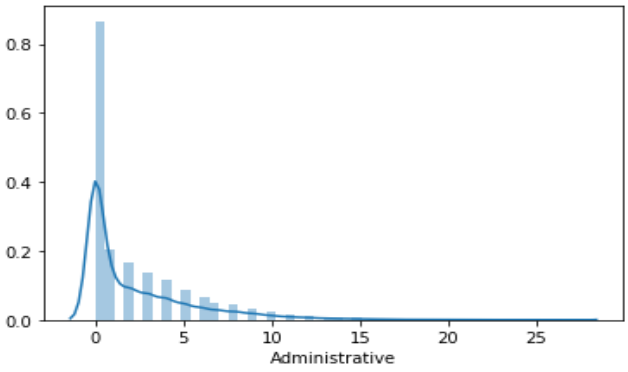
CHECKING OUTLIERS FOR NUMERICAL VARIABLES

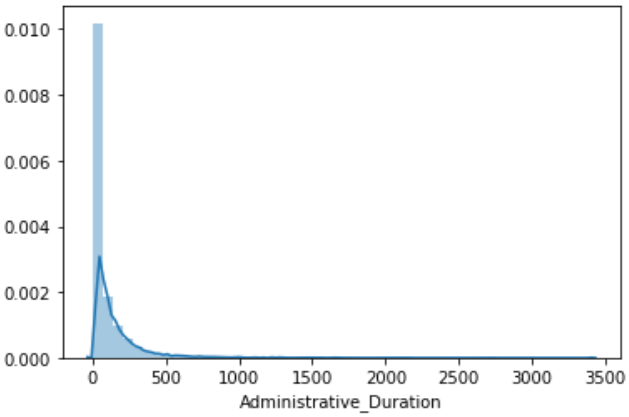


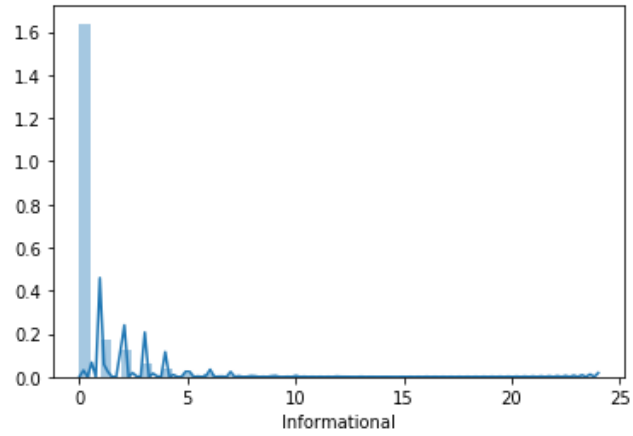
As we have outliers in many of the attributes we treat them using square roottransformation.

BY USING SQUARE ROOT TRANSFORMATION

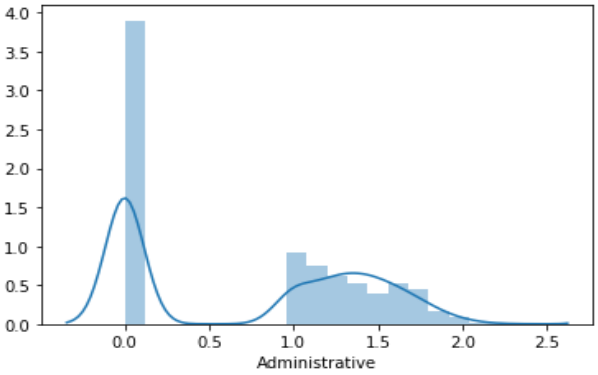
Before

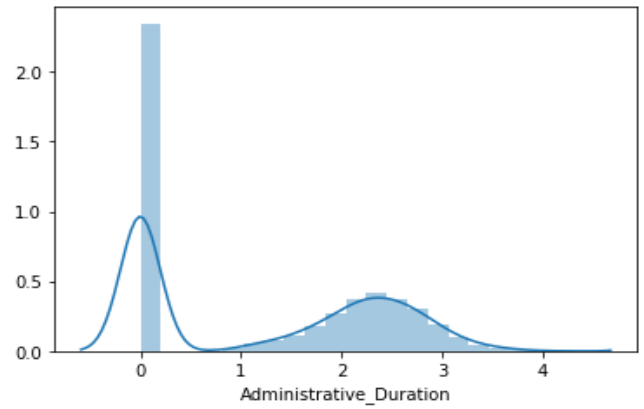


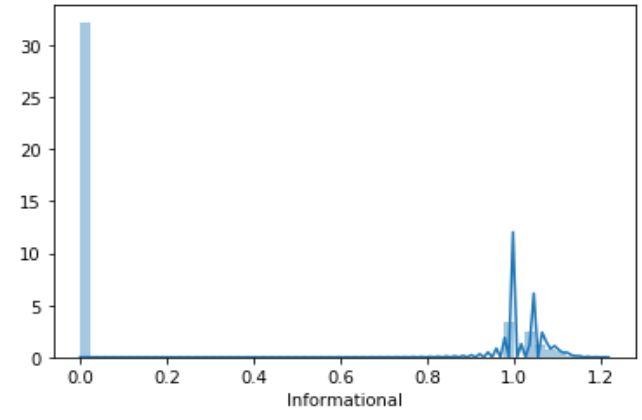


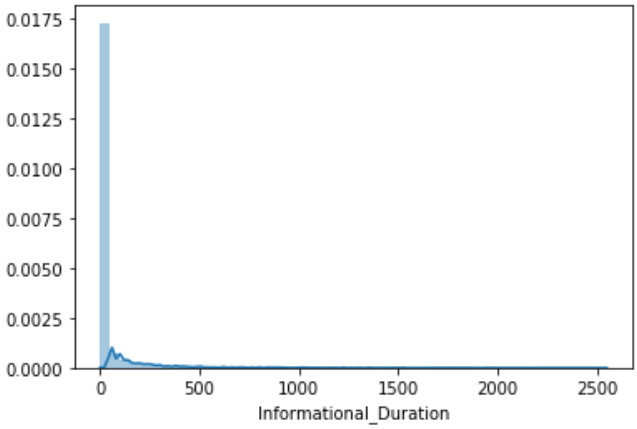


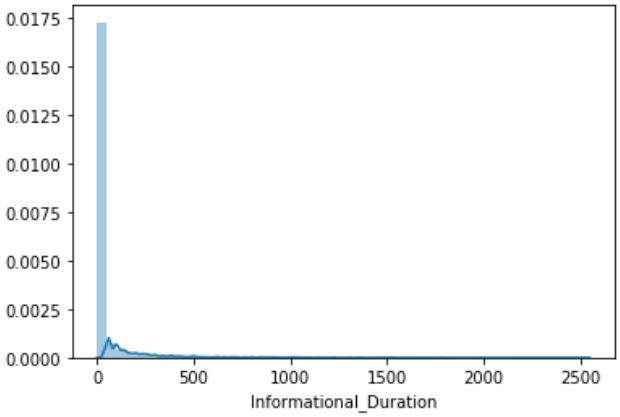
After

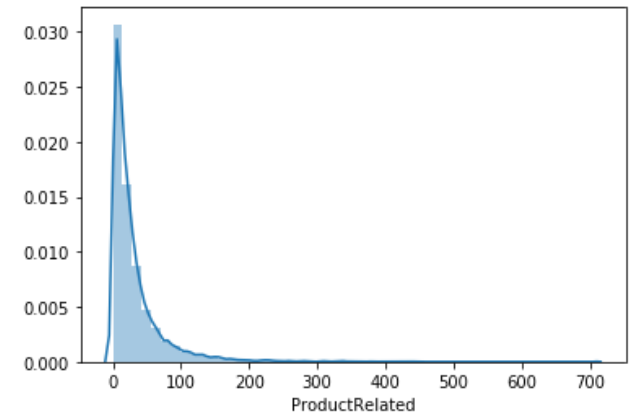


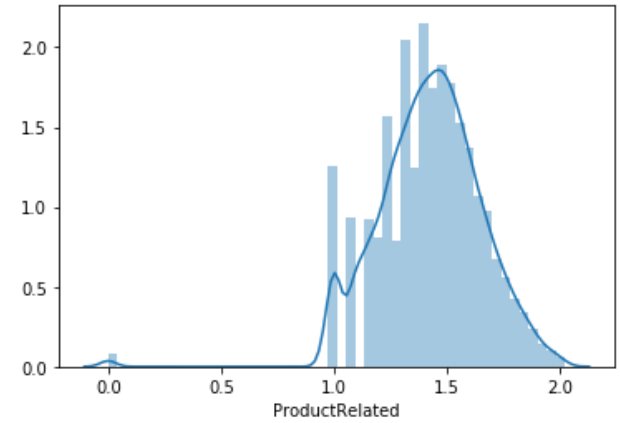


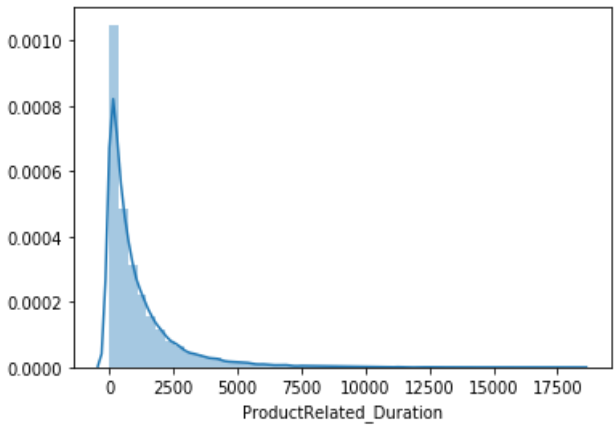


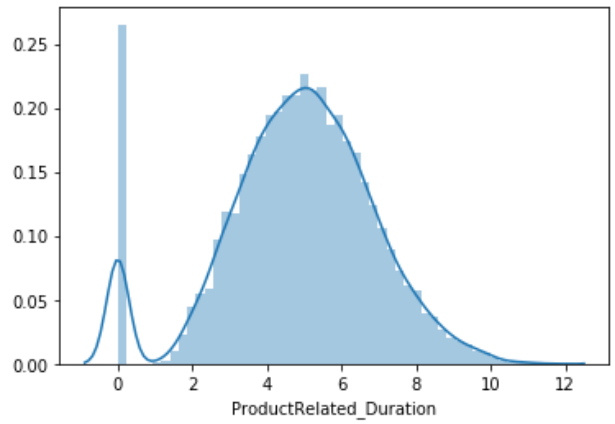

0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
500
1000
1500
2000
2500
Informational_Duration

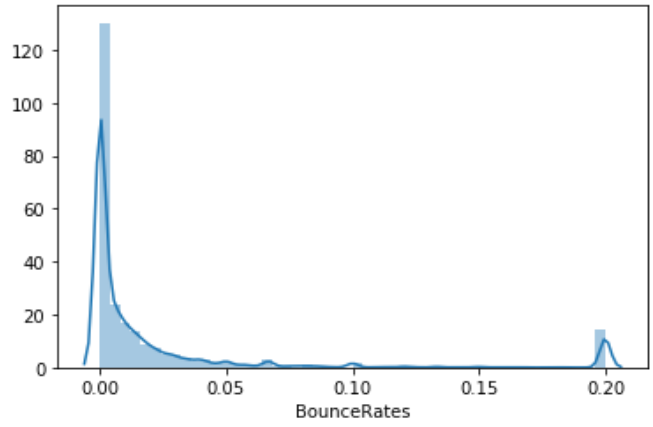

0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
500
1000
1500
2000
2500
Informational_Duration

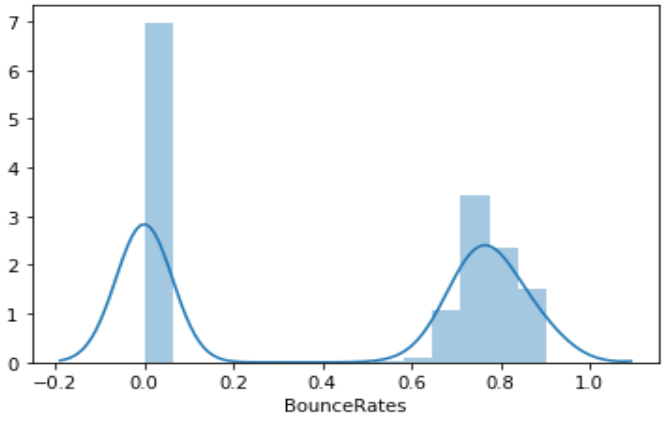

0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500
600
700
ProductRelated

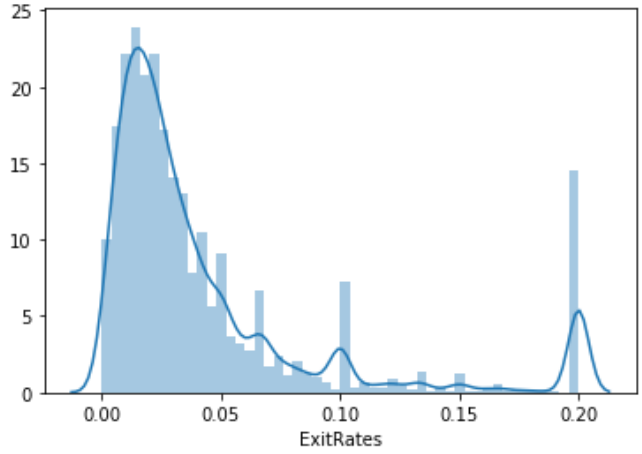

2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
ProductRelated

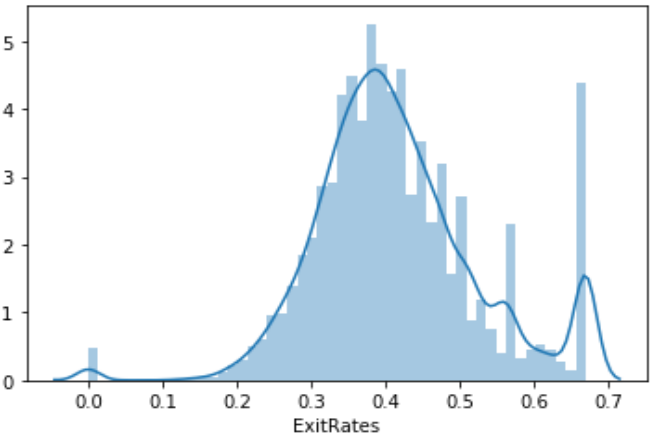

0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
2500
5000
7500
10000
12500
15000
17500
ProductRelated_Duration

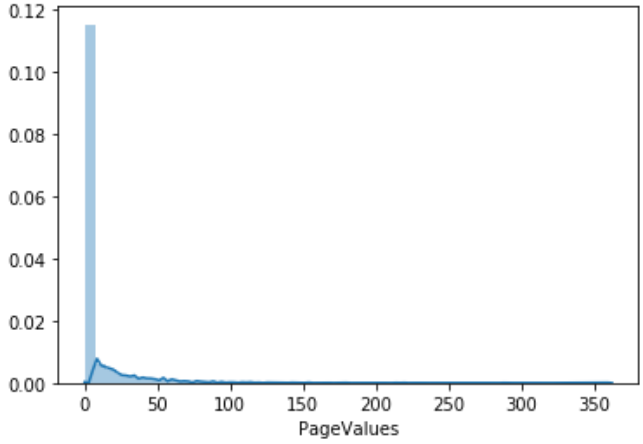

0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
ProductRelated_Duration

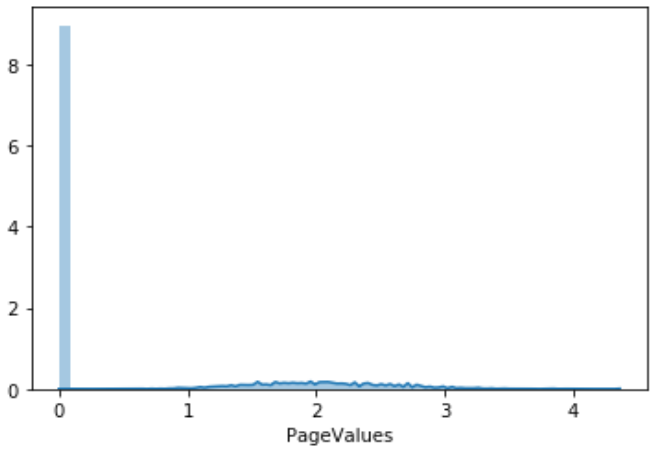

BounceRates


BounceRates


ExitRates


ExitRates


PageValues


PageValues

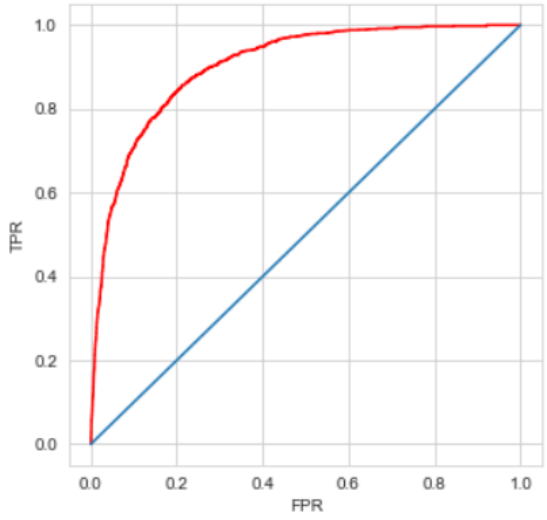
CHAPTER-5

MODELLING (Basic Model)

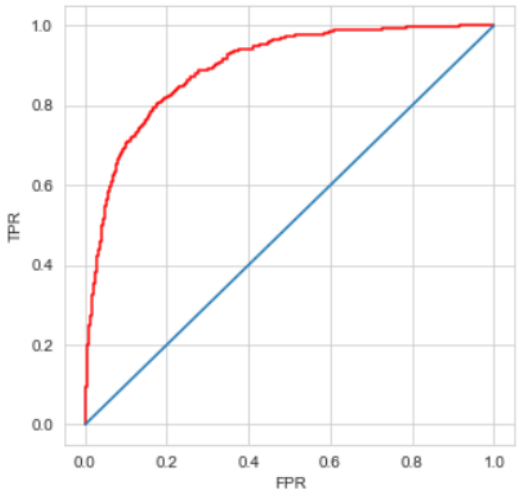
Logistic regression is the extension of the linear regression where the target variable is categorical in nature and not continuous. It predicts the probability of the outcome variable.

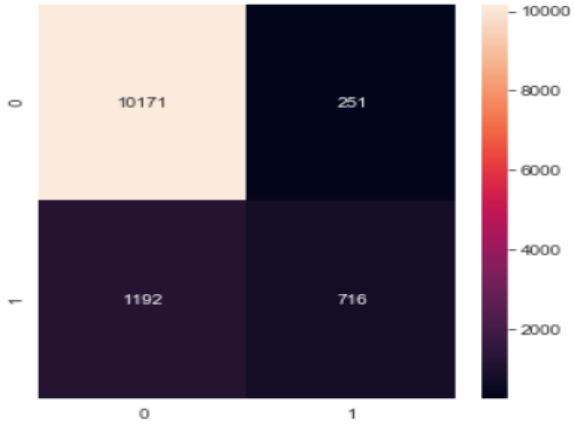
Logistic Regression using all the variables

```
AUC score of Train: 0.9027355838685623
[[7114  184]
 [ 821  512]]
```




```
AUC score of Test: 0.8952090408061014
[[3054   70]
 [ 372  203]]
```





Precision and Recall:

	precision	recall	f1-score	support
0	0.90	0.98	0.93	10422
1	0.74	0.38	0.50	1908
accuracy			0.88	12330
macro avg	0.82	0.68	0.72	12330
weighted avg	0.87	0.88	0.87	12330

OTHER MODELS

We will compare five different machine learning models using the great Scikit-Learn library:

- **Logistic Regression:**
  Logistic regression is the extension of the linear regression where the target variable is categorical in nature and not continuous. It predicts the probability of the outcome variable.

  Pros:

  1. Don't need to pick learning rate
  2. Often run faster (not always the case)
  3. Can numerically approximate gradient for you (doesn't always work out well)

  Cons:

  1. More complex
  2. More of a black box unless you learn the specifics

- **Naive-Bayes:**
  It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

  Pros:

1. Easy to implement
2. Requires a small amount of training data to estimate the parameters
3. Good results obtained in most of the cases
4. Handle Missing Values by ignoring the instance during the probability estimate calculations

   Cons:

1. Assumes Independence of features
2. Independence existence may not hold for some attributes

3. Practically Dependencies exist among the variables
4. So loss of accuracy due to these reasons

- **K-Nearest Neighbors Classifier:**
  K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data.

  Pros:

- It is extremely easy to implement
- It is lazy learning algorithm and therefore requires no training prior to making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g SVM, linear regression, etc.
- Since the algorithm requires no training before making predictions, new data can be added seamlessly.
- There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

  Cons:

- The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate distance in each dimension.
- The KNN algorithm has a high prediction cost for large datasets. This is because in large datasets the cost of calculating distance between new point and each existing point becomes higher.
- The KNN algorithm doesn't work well with categorical features since it is difficult to find the distance between dimensions with categorical features.

- **Decision Tree Classifier:**
  Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Pros:

1. It is able to generate understandable rules.
2. It performs classification without requiring much computation.
3. It is able to handle both continuous and categorical variables.
4. It provides a clear indication of which fields are most important for prediction or classification.

Cons:

1. It is less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
2. It is prone to errors in classification problems with many class and relatively small number of training examples.
3. It can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

- **Random Forest Classifier:**
  A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

  Pros:

1. The predictive performance can compete with the best supervised learning algorithms
2. They provide a reliable feature importance estimate
3. They offer efficient estimates of the test error without incurring the cost of repeated model training associated with cross-validation

Cons:

1. An ensemble model is inherently less interpretable than an individual decision tree
2. Training a large number of deep trees can have high computational costs (but can be parallelized) and use a lot of memory
3. Predictions are slower.

To compare the models, we are going to be mostly using the Scikit-Learn defaults for the model hyper parameters. Generally these will perform decently, but should be optimized before actually using a model. At first, we just want to determine the baseline performance of each model, and then we can select the best performing model for further optimization using hyper parameter tuning.