Analyzing Online Shoppers' Intention

Date: 01st June 2019

Contact : +61466639672

Author :

Gururaj Gosala – s3774180 (MC267)

Aman Devalla – s3752307 (MC267)

Table of Contents :-

Abstract:

Looking back to 2015, less than 10% of Australians used a smartphone to make an online purchase, Fast forward to 2017, and one in five online purchases were made from a mobile device. The current dataset Online Shoppers' Intention gives a view at the factors that are crucial and tend to play a major role in the business strategies of the companies that mostly work over the online platform. Our Goal in this assignment is to "**Identify key metrics that impact a shopper's revenue-conversion on the site**". We have inspected the relationships between the attributes that influence the online shoppers' behaviour and based on that we have tried to change the aspect by which there could be an improvement in the output as per the goal of our assignment.

Introduction:

- The dataset consists of 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.

The dataset consists of 10 numerical and 8 categorical attributes.
The 'Revenue' attribute can be used as the class label.
The information of the dataset is given below, all the columns are mentioned as per respective order in the dataset and explained in a proper manner.

- "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year. We have chosen "Online Shoppers Purchasing Intention Dataset" from UCI ML Repository (source: as per the mention in assignment)

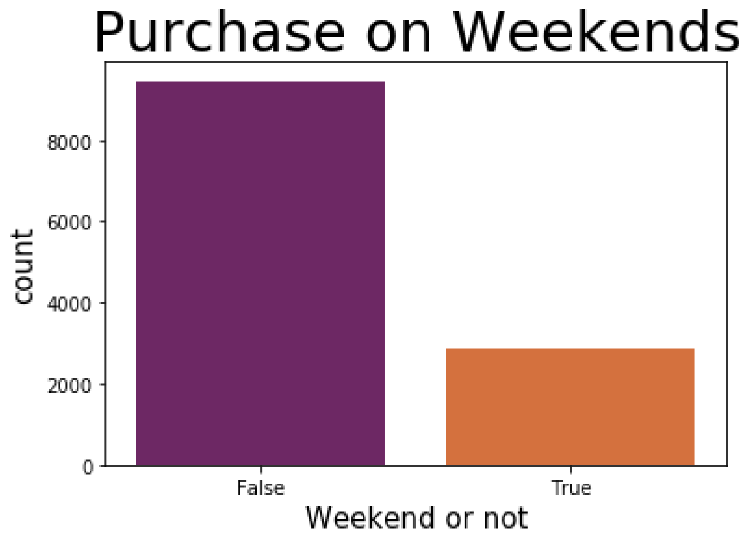
- The data consisted of multiple columns as mentioned above.
- We have plotted frequency plots of multiple fields like Revenue, Weekend, Informational, Administrative, etc. and also for attribute pairs like 'Revenue – Weekend', 'Revenue – Region', 'Revenue – Informational', etc. to understand the relationship of each attribute and also to understand the business use-cases of these respective attributes in maximizing the revenue-conversion on the e-commerce site

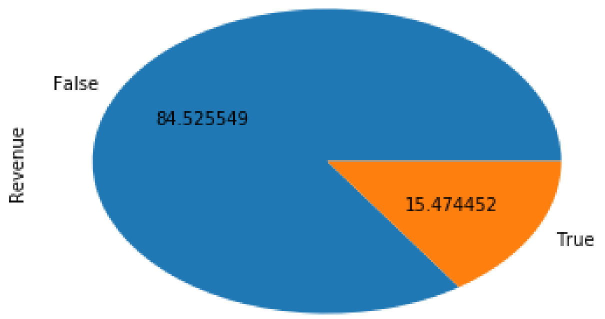
We have kept in mind of the goal we had to get to and our intention was clearly to make out the impacting metrics and try and make a better analyzation in order to improve the methodology of the revenue conversion based on the data provided.

Methodology:-

In the beginning we have depicted the relations between every two attributes and cross-verified if they would affect the revenue as we are trying to improve the revenue by targeting few key metrics. The output between different attributes has showed us where the issues could be and how we could change the system in order to improve the potential visitors from skimming through portal to purchasing a product they are looking for.

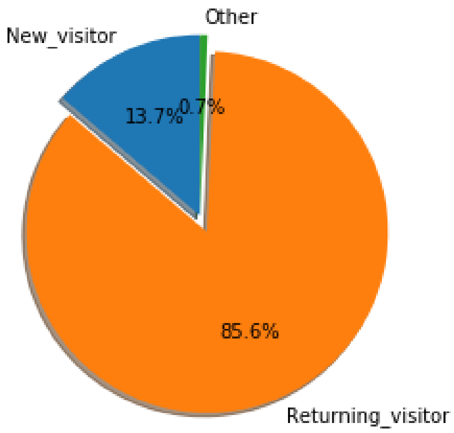
To implement this the understanding of the relations between the metrics is important as that is the key for getting to the depth. Few of the relations we have depicted are:





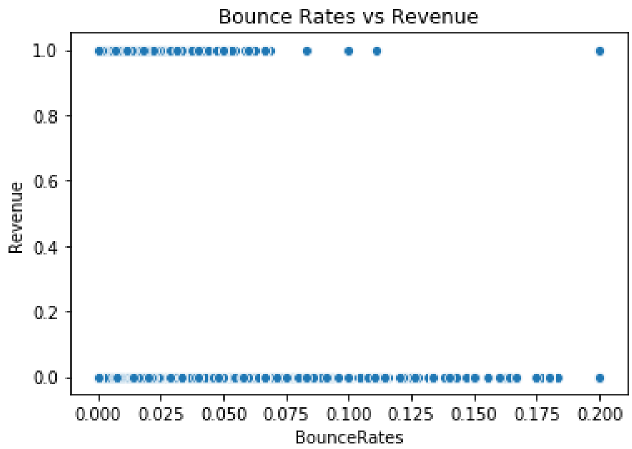
The Overall revenue percentage, one depicts in terms of weekends and the other depicts the revenue % of the total dataset where True is the number of revenue conversions successful.

This shows us that there are many people who are not influenced by what they are seeing over their devices and therefore are not interested in buying any product.

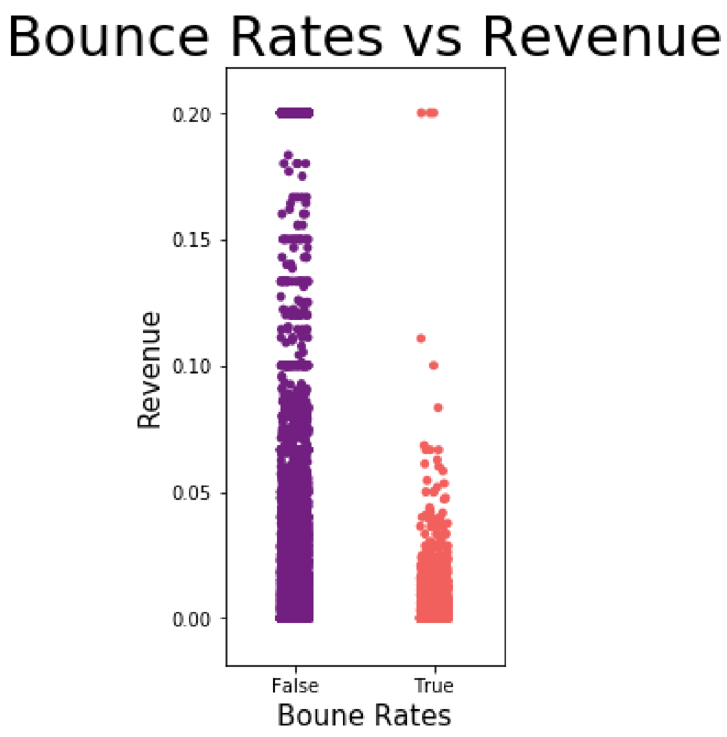


The above image gives us the statistical representation of new visitors percentage and that is always a good platform when new visitors tend to visit a respective online shopping portal. They tend to be searching for a product which in case provided with proper recommendation can be a successful revenue converter.

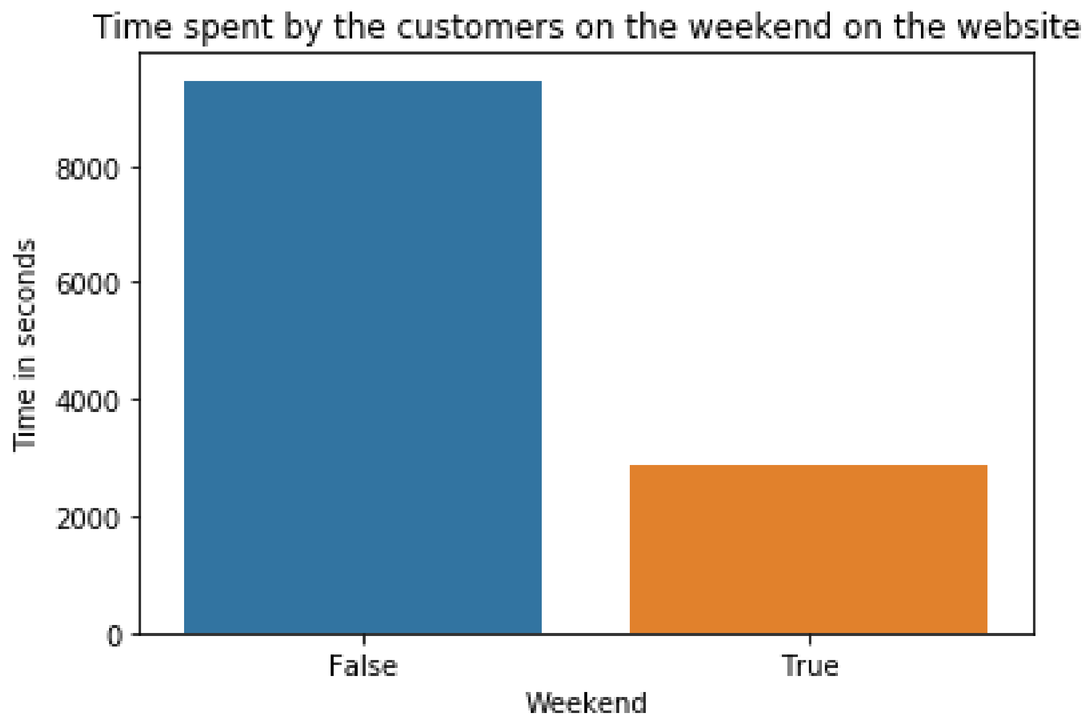
By checking through the dataset as to how many number of observations have been recorded in terms of months and the months with most recorded observations are May, November, March. As May has Mother's Day most of the people tend to buy gifts for their Mother. These months could be potential revenue converters and must be properly strategized in order to find out what could be the key metric to target in order to convert the potential customers to revenue converters.



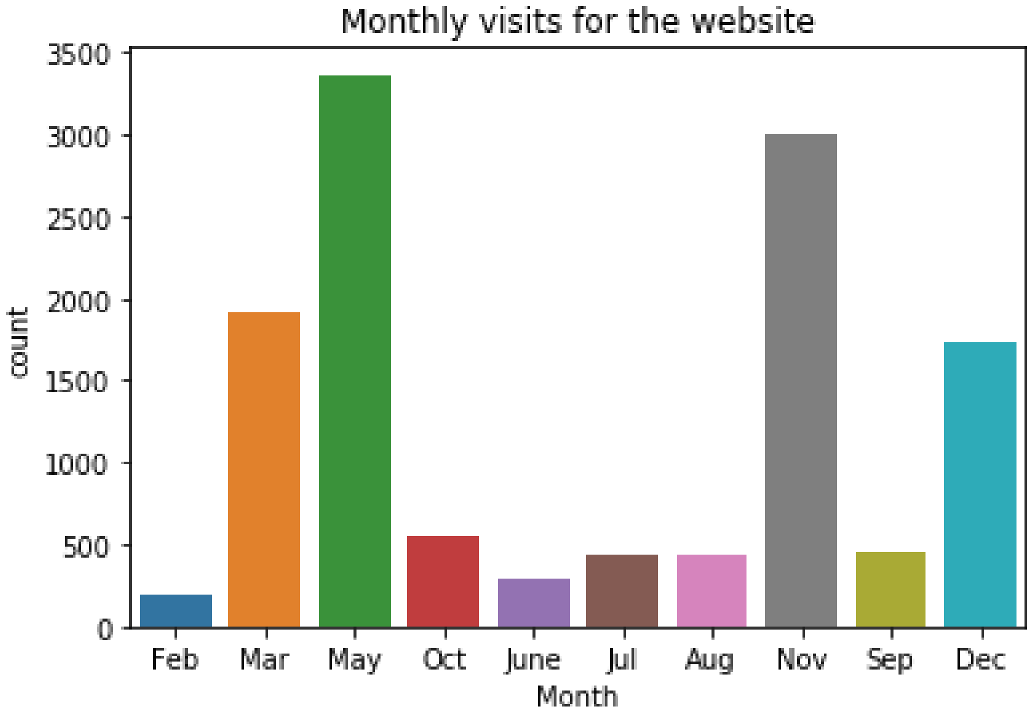
Bounce rate have a definition which tend to explain the interest every customer has depending on what they see on their online portal and it is a crucial attribute as the percentage of bounce rate when higher implies that people are less interested in the products they are recommended and thus it has to improve the accuracy of what the customer or consumer requires. Revenue could be directly proportional to the least bounce rate. If the prediction and recommender algorithm is successful most of the time then the customer tends to like the product and therefore spend more time on the online portal thus can become a revenue converter which tends to be the main aspect of this assignment.



Bounce rates compared to revenue shows the impact of customers. The bounce rates are typically high compared describing that there was no revenue generated.



When we compare the weekend data’s we can see that the time spent was also not high on the weekends and it can be improved.



If we have a look at the page visits, we can see a maximum of visits in month May and November stating that there are more people interested in shopping online in those months.

We process the data, then train and test the data by split. We have used 70:30 split which showed the maximum efficiency. We found the training accuracy to be 99% efficient and testing accuracy to be 88%. We have used Random Forest method and it was effective to provide the results.

Results:

The revenue generated being the dependent variable can be seen to be very low in the analysis. It can be improved by lowering the bounce rates, as the bounce rates declare the time spent to close the page, which is high in this case. The bounce rates can be increased by analyzing the customer's behavior on different products and their interests. In addition, the investigation shows that the revenue did not get a huge impact even on the special days. The special days had the number of customers logging in but the revenue could not be generated. The factors such as bounce rates, page values and Weekends need to be considered in order to get customers to generate revenue. In addition, when we consider the data, we see there are many returning customers with very less revenue generated, so new products can be added to match the interests of the returning customers to increase the revenue using the keywords searched.

Discussion:

The data is collected over a time of 1 year and it can be too early to make any kind of predictions and also this being a commercial market, it is a hard domain to excel as it depends on the various products and various interests of the customers and this can only be achieved by getting people more engaged by introducing the products that can interest them and looking for the trends that people are looking for.

References:

https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset