Online Shopper's Purchasing Intention

Project By: Aishwarya Sanganalu Mattha

1. Overview

The objective of this project is to create a logistic model used to determine the probability of a customer making a transaction on an eCommerce website. This model will help a company in determining their current customer base and persuade potential customers with personalized marketing, hence increase in company's performance.

2. Data Exploration

```
> 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
 $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE
 $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE
```

Figure 1: Structure of the dataset

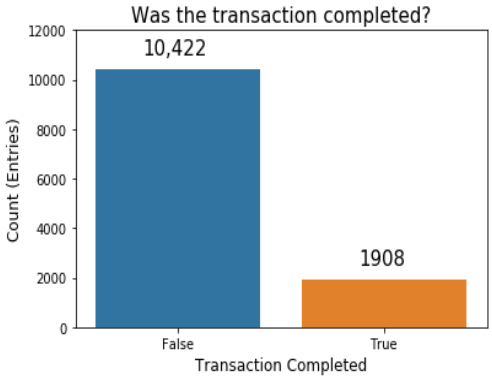

Figure 2: Number of observations in each category of purchase

The dataset contains 18 variables: 10 numerical and 8 categorical variables. This dataset has 12330 entries, split into 10,422 entries where the shoppers did not purchase and 1908 entries where the shoppers did purchase (Figure 2). Each entry is based on unique users in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

For preprocessing, the dataset does not contain null values but some of the features needs to be converted to factors. A One Hot Encoding technique can be used to encode our string variables to integer labels, then convert our labels from integers to One Hot columns to remove any implied hierarchy.

Because the data is so heavily skewed in the direction of the 'No purchase made' category, the most important step is to reduce the class imbalance for the predicting variable 'Revenue'. With Synthetic Minority Over-sampling Technique (SMOTE) number of minority instances will be increased by creating synthetic datapoints from the most influential variable in the dataset.

2.1 Impact of Bounce Rate and Exit Rate

Bounce rate is the overall percentage of a single engagement session whereas exit rate is the percentage of exits from a page. Hence the former is calculated by dividing the aggregation of one-page visits to the overall entrance visits whereas latter is calculated by dividing the aggregation of total exits from a page to the total visits to a page. One major difference between these closely tied metrics is that exit rate is related to the overall percentage of visitors that were within the last session whereas bounce rates account for the percentage of visitors that were part of that one and only session. Hence in the case of bounce rate, prior activity is not considered. Hence all bounces logically define exits but conversely it is not true.

Figure 3 shows a high correlation between 'ExitRates' and 'BounceRates'. A high bounce rate could indicate issues with user satisfaction [1] owing to one or many reasons such as unfriendly UI of the website, extremely slow throughput or other technical matters. A high exit rate could be a sign of lower performing sectors in funnels, showing areas open to optimization as if customers are leaving then at the end of the day no one is buying. According to BigCommerce [2], a bounce rate between 30% to 55% is acceptable. Our analysis shows the bounce rates largely scattered lower than 10% (Figure 4). Which means many of the customers are leaving the website as soon as they land to the eCommerce webpage.

For the other variables, the correlation plot (Figure 3) shows that most of the numerical attributes seem to exhibit high positive skewness whereas some exhibit nominal tinge of negative skewness.

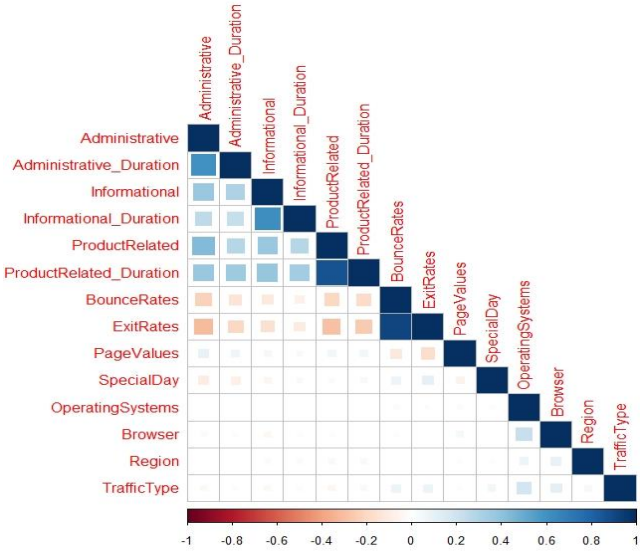

Figure 3: Correlation plot of numeric variables in the dataset

Figure 4: Relationship between Exit rates and Bounce rates

2.2 Impact of loyal customers and "weekend syndrome"

The Figure 5 shows that most of the customers whether they drive in revenue or not, are returning customers. This analysis shows that the eCommerce company is doing good with respect to keeping their current customers satisfied. Also, most of the visitors came in and made a purchase during the weekday (Figure 6).

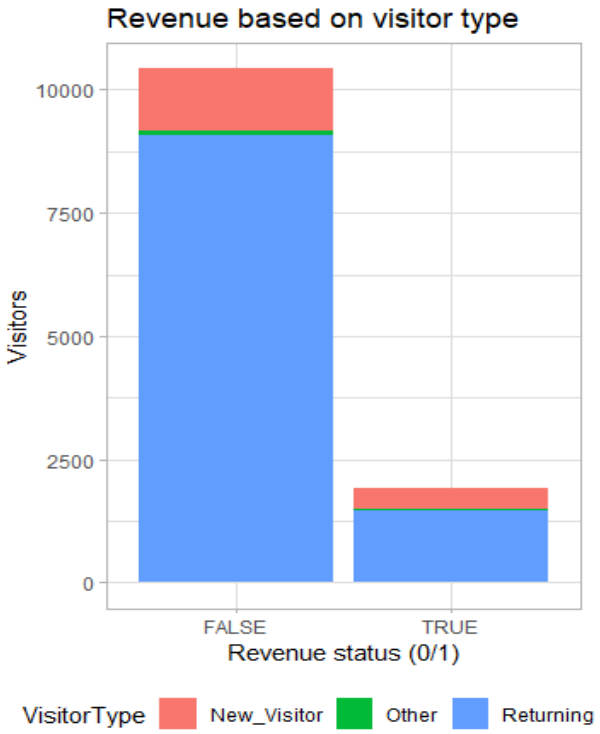

Figure 5: Number of visitors making a purchase vs no-purchse

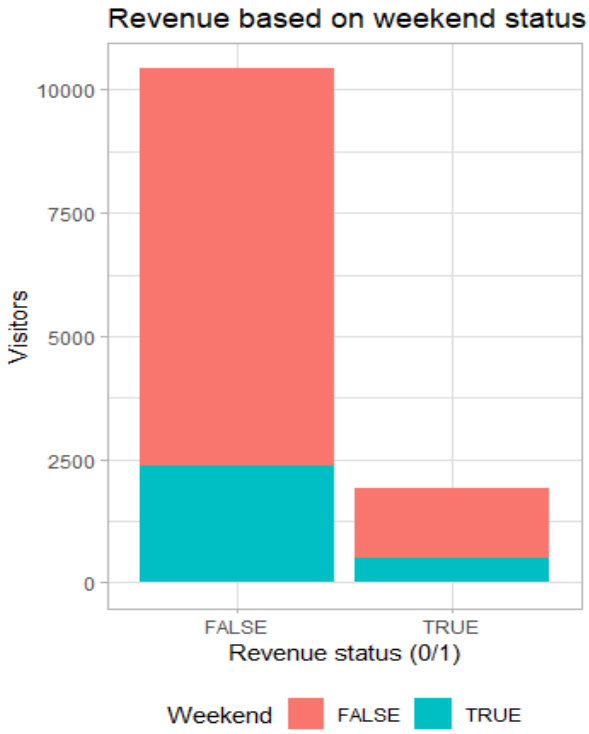

Figure 6: Sales on weekend and weekday

2.3 Revenue during holidays

As we all know, most of the purchases happens during holidays. Seeing this pattern indicates the true nature of the shoppers reflecting in our dataset. The following figure (Figure 7) depicts the seasonality revenue growth. There seems to be a high customer engagement during the months of Mar and May, post which the trend seems to be decreasing. Moreover, between the months of June to Oct the trend seems to stagnate post which there seems to be high engagement as Black Friday, Christmas, and New Year approaches. When the demand appears high, there appears to be a lot of engagement but significantly lower conversion rates as most of these purchases are driven by returning customers (Figure 8).

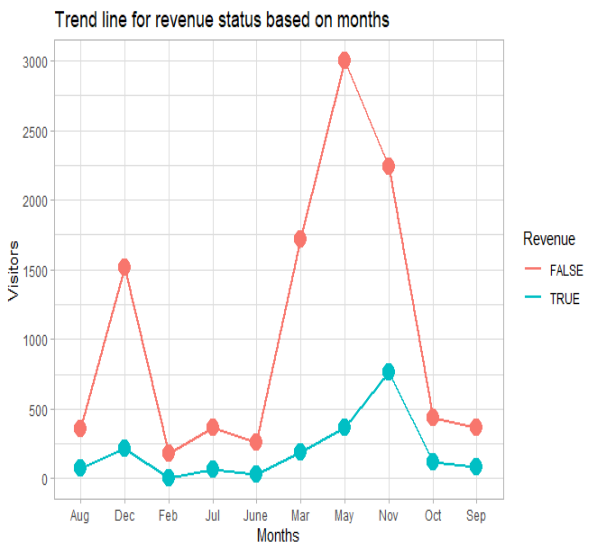

Figure 7: Purchases made on each month

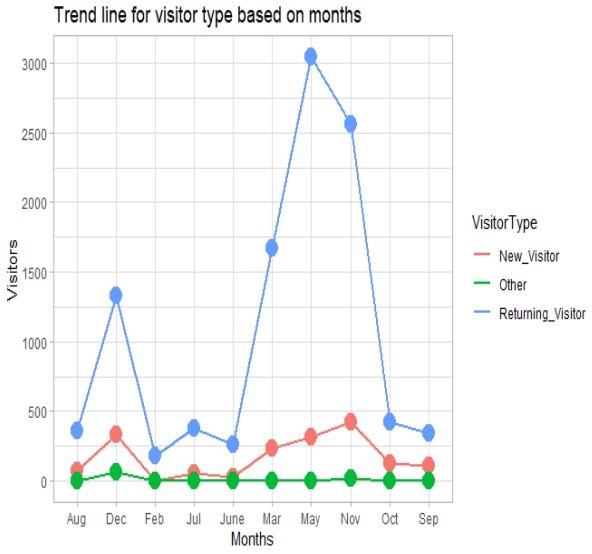

Figure 8: Relating to Visitor Types who are potential purchasers

2.4 Other Revenue drivers

From Figure 9, we can capture the relationship between revenue growth and the operating system. The top performer remained "2" in both cases i.e., visitors who did not made a purchase and visitors who made a purchase. However, following positions were conversely secured by "1" and "3". Other sources brought in considerably lower customers. This could either mean that the website is not user friendly on those sources or simply because those sources are niche, not many customers use them.

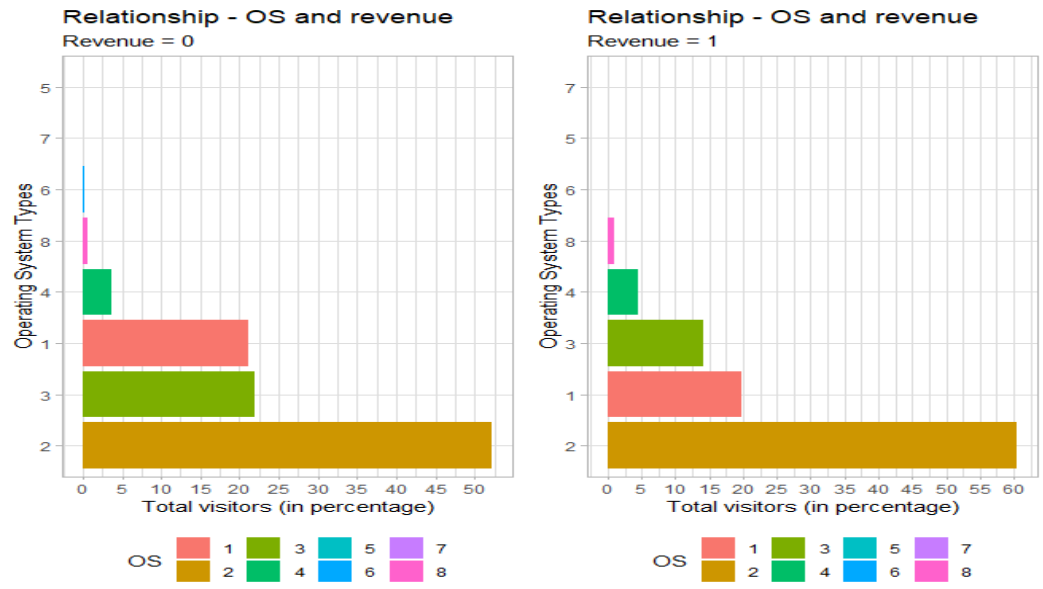

Figure 9: Relationship between OS and Revenue

Figure 10 shows browser “2” remains at the top followed by “1”, “4” and “5” in both cases. This could suggest the same reasonings as OS. With respect to region (Figure 11), “1” seems to be performing significantly better followed by “3” in both cases. The lead of “1” is highly significant suggesting that marketing reach within this region is well versed with.

Figure 10: Relationship between Browser and Revenue

Figure 11: Relationship between Region and Revenue

3. Conclusion

There are groups of customers where majority of them follow a same purchasing pattern. Most of them prefer to buy on weekdays than on weekends. Also, most of the customer population use browser “2”, so marketing on browser “2” will reach many of the new population. With most purchases being made in region “1”, we can help establish a new upcoming company of similar market in region “1”. A logistic model can then predict how the market would perform outside of the current establishment. This can help a company grow and to target potential customers with personalized ads.

References

[1] *CXL*. (n.d.). Retrieved from Bounce Rate vs. Exit Rate: What's the Difference?: https://cxl.com/guides/bounce-rate/bounce-rate-vs-exit-rate/

[2] Moser, J. (n.d.). *BigCommerce*. Retrieved from How Personalization Can Reduce Ecommerce Bounce Rates by 20-30%: https://www.bigcommerce.com/blog/bounce-rates/#what-is-a-bounce-rate-on-an-ecommerce-website