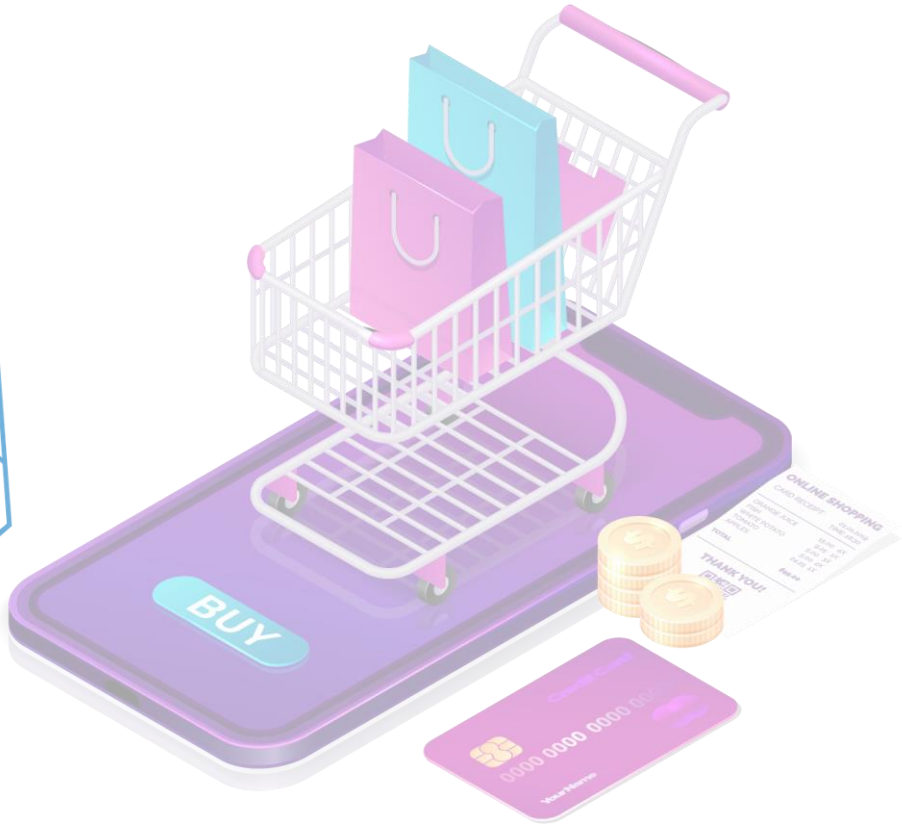
GREAT LAKES INSTITUTE OF MANAGEMENT

greatlearning

PGPDSE-FT.G.JUL19 Batch

Capstone – Project Report

Online Shoppers Purchasing Intention



Submitted By:

Surbhi Bhatia

Siddharth Shankar

Rahul Das

Vaibhav Bhatnagar

Jalaj Tripathi

Mentored By:

Mr. Romil Gupta

ABSTRACT

Consumers shopping activities on the internet turn out to be more important every year. Although the increase of e-commerce usage over the last few years has created potential in the market, most of the visitors still do not complete their online shopping process. Almost 95 percentage of internet users visit online retailers without purpose of actually making a transaction.

Even when consumers visit online retailers in purpose of making a purchase, many of them do not finish the transaction and abandon their purpose just prior to checkout. According to research, the average rate of consumers who do not fulfil the process of shopping transaction is approximately 70 percent in 2010.

This leads the online retailers the need for solutions to prevent the loss of their revenues.

Problem Statement: Predict the purchasing intention of the visitor using aggregated pageview data track during the visit along with some session and user information

The aim of this study is to evaluate the actions taken by the visitors on ecommerce environment in real time and predicting the visitor's shopping intent.

The objective of this project is to track the mouse movements, the link and button click information that the user has on the screen and the tracking data of the pages visited will be obtained and the actions taken as the result of these data will be determined. Acceptable actions will be used as labels during pattern definition with supervised learning algorithms. Thus, when any user receives actions that match the predefined pattern, they will be tagged with the obtained pattern function and the action to be taken instantaneously will be determined.

The extracted features from page view data kept track during the visit along with some session and user information are fed to machine learning classification methods to build a model.

ACKNOWLEDGEMENT

We hereby certify that the work done by us for the implementation and completion of this project is absolutely original and to the best of our knowledge. It is a team effort and each of the member has equally contributed in the project.

Date: 20-11-2019

Place: Gurugram

CERTIFICATE OF COMPLETION

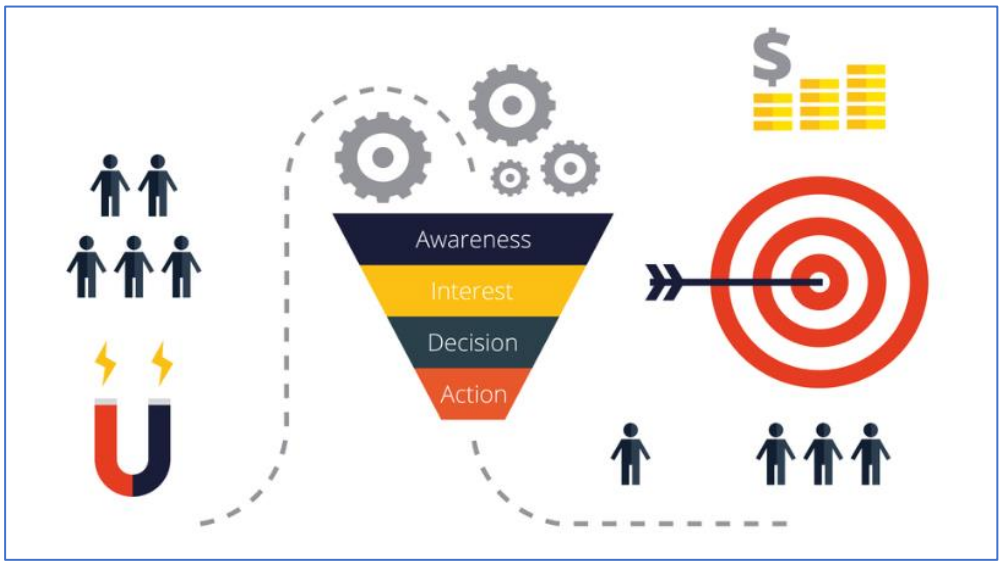
This is to certify that the project titled **"CAPSTONE – PROJECT REPORT – ONLINE SHOPPERS PURCHASING INTENTION"** was undertaken and completed under the supervision of Mr. Romil Gupta for the Post Graduate Program in Data Science and Engineering (PGPDSE).

Mentor: Mr. Romil Gupta

EXECUTIVE SUMMARY

Background and Need:

According to Forbes, Ecommerce will become the largest ***retail channel*** in the world by 2021[1] and to cater such a major increase in the online retail and ecommerce activity retailers will need to tailor their digital strategies appropriately. One such aspect is to predict the intent of the online shoppers in real time and present them with offers if they are likely to abandon the site. The purchase conversion rate is an important factor demonstrating how successful the website is and the digital strategies of the company.



Source: digitalmarketinginstitute.com

Scope and Objective:

The scope of the project is to predict the purchasing intention of the visitor using aggregated pageview data track during the visit along with some session and user information.

The objective of this project is to track the mouse movements, the link and button click information that the user has on the screen and the tracking data of the pages visited will be obtained and the actions taken as the result of these data will be determined. Acceptable actions will be used as labels during pattern definition with supervised learning algorithms. Thus, when any user receives actions that match the predefined pattern, they will be tagged with the obtained pattern function and the action to be taken instantaneously will be determined. The visitor behaviour analysis model is designed as a binary classification problem measuring the user's intention to finalize the transaction. In order to predict the purchasing intention of the visitor using aggregated page view data kept track during the visit along with some session and user information are fed to machine learning classification methods to build a model.

Approach and Methodology:

The dataset has been downloaded from the UCI Machine Learning Repository. It is to be noted that the available dataset is a subset of the original data as found in the research paper.[2] After structuring, cleaning and outlier treatment, oversampling and feature selection pre-processing steps are used to enhance the performance and scalability of the classification methods. Various machine learning methods were used under different

[1] http://forbes.com/sites/michellegrant/2018/08/14/e-commerce-set-for-global-domination/
[2] https://doi.org/10.1007/s00521-018-3523-0

conditions to check the performance and scalability. Further, the models are evaluated using relevant performance metrics and the robust model was selected.

PROJECT OVERVIEW

The online marketing space is in the constant shift as new technologies, services, and marketing tactics gain popularity and become the new standard. Online store owners are one of the many different segments affected by these constant evolutions and competitors in the online retail space. In order for these business owners to survive and thrive, they need to be able to make better decisions faster. This is where web analytics comes into play for ecommerce sector. The *online shoppers purchasing intention dataset* provides some of the metrics from the 'Google Analytics' combined with user click information.

The dataset has been collected from www.columbia.com.tr which was provided by the Turkish Gözalan Group. The Columbia Sportswear Company is an American company that manufactures and distributes outerwear, sportswear, and footwear, as well as headgear, camping equipment, ski apparel, and outerwear accessories. It was founded in 1938 (81 years ago)[3]. In Turkey, the Gözalan Group became the distributer of Columbia Sportswear in 2004[4] and according to the Turkish website of Columbia, it was registered in the year 2008. No information is available about the year of the data extracted. So, we assume that the dataset is either of 2009 or later.

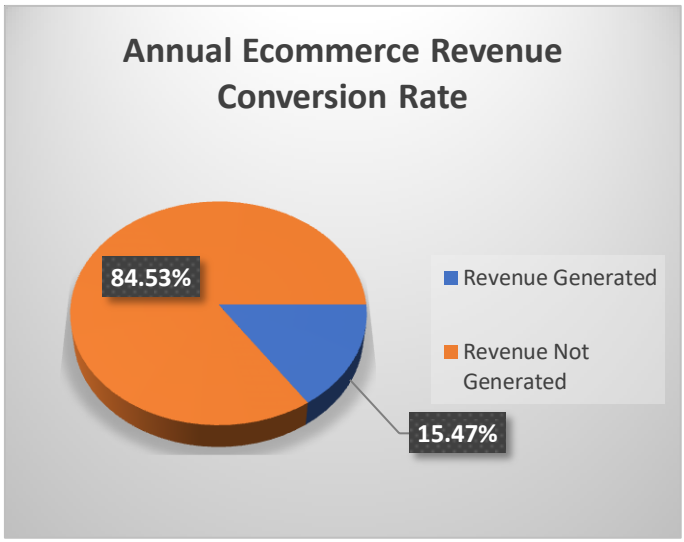


Fig: Current Conversion Rate

The dataset consists of feature vectors belonging to 12,330 sessions. Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest 15.5% (1908) were positive class samples ending with shopping.

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. Use the dataset to build a Machine Learning Model to predict whether the user visit has been finalized with a transaction.

Primarily the dataset consists of 10 numerical and 8 categorical attributes. *Revenue is the class label (target).*

Data Description:

Numerical Features:

Feature Name	Feature Description	Min Value	Max Value	Std. Dev.
Administrative	Number of different types of pages visited by the visitor about account management	0	27	3.32
Administrative Duration	Total amount of time (in seconds) spent by the visitor on account management related pages	0	3398.75	176.78

[3] https://en.wikipedia.org/wiki/Columbia_Sportswear
[4] https://www.nebim.com.tr/en/columbia

Informational	Number of different types of pages visited by the visitor about website, communication and address information of the shopping site	0	24	1.27
Informational Duration	Total amount of time (in seconds) spent by the visitor on informational pages	0	2549.37	140.75
Product Related	Number of different types of pages visited by visitor about product related pages	0	705	44.47
Product Related Duration	Total amount of time (in seconds) spent by the visitor on product related pages	0	63973.52	1913.67
Bounce Rate	Average bounce rate value of the pages visited by the visitor	0	0.2	0.04
Exit Rate	Average exit rate value of the pages visited by the visitor	0	0.2	0.05
Page Value	Average page value of the pages visited by the visitor	0	361.76	18.57
Special Day	Closeness of the site visiting time to a special day	0	1	0.19

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor and total time spent in each of these page types in seconds. 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.

For example:

Administrative clicks can be – login, logout, password recovery, profile, email wish list
Informational clicks can be – ad popups, contact us, nearby stores
Product Related clicks can be – search, shopping cart

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. These values can be stored in the system for all web pages of the e-commerce site in the developed system and updated automatically at regular intervals.

"Bounce Rate" - Value 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.

"Exit Rate" - Value is calculated as for all page views to the page, the percentage that were the last in the session.

"Page Value" - 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 the special days (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.

Categorical Features:

Feature Name	Feature Description	Number of Categorical Values
Operating Systems	Operating system of the visitor	8
Browser	Browser of the visitor	13
Region	Geographic region from which the session has been started by the visitor	9
Traffic Type	Traffic source by which the visitor has arrived at the website (e.g. banner, SMS, direct)	20
Visitor Type	Visitor type as “New Visitor”, “Returning Visitor” and “Other”	3
Weekend	Boolean value indicating whether the date of the visit is weekend	2
Month	Month value of the visit date	10
Revenue	Class label indicating whether the visit has been finalized with a transaction	2

Data Cleaning and Formatting:

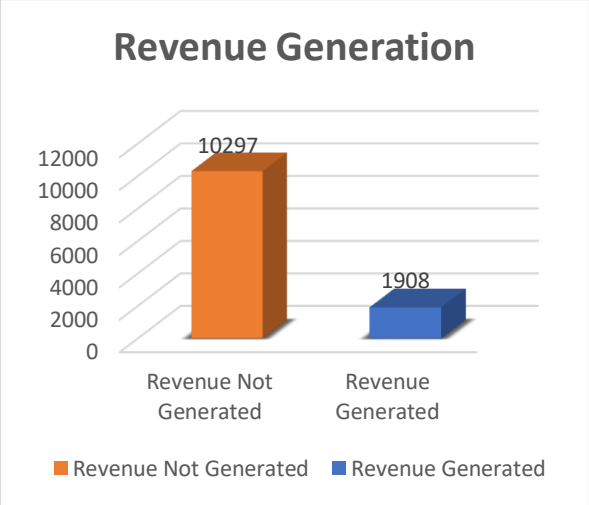
Though it has been mentioned that the dataset was so formed that each session would belong to a different user in the in a 1-year period but there are 125 duplicate records, so we dropped them. Now, the current shape of the dataset is (12205, 18). There are no missing values.

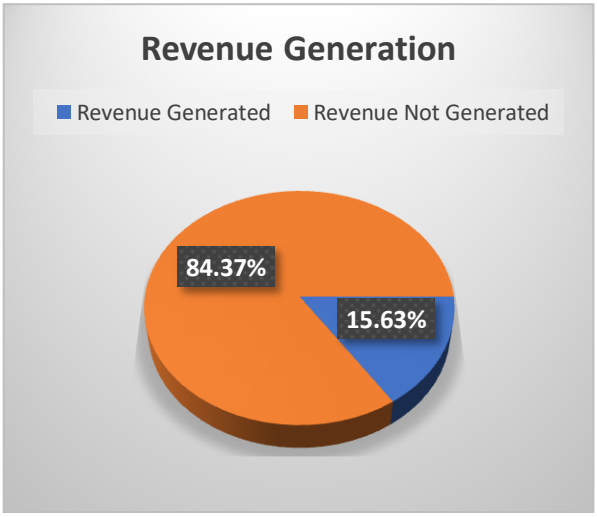
Missing Values	There are no missing values.	
Duplicate Records	There are 125 duplicate records.	
Outlier Treatment	Distribution plots were made to check the skewness and outliers in the dataset.	Since, outliers are an important part of the dataset we retained them and transformed to map them to normal distribution.
Encoding	Operating System, Browser, Region, Traffic Type, Visitor Type, Special Day and Weekend were One-Hot Encoded.	Month and Revenue features were Label Encoded.
Standardization	We apply Robust Scaler to scale features using statistics that are robust to outliers.	
Sampling	We applied both under-sampling and oversampling methods and compared their effects.	

EXPLORATORY DATA ANALYSIS

The purpose of EDA is to find anomalies, patterns, trends, or relationships in the given dataset. To begin the EDA, we will focus on a single variable, the Revenue, because this is the target for our machine learning models.

There are 1,908 sessions or customers out of 12,205 who have made a transaction. The baseline revenue generation through Online Shoppers for the given year is **15.63%**.

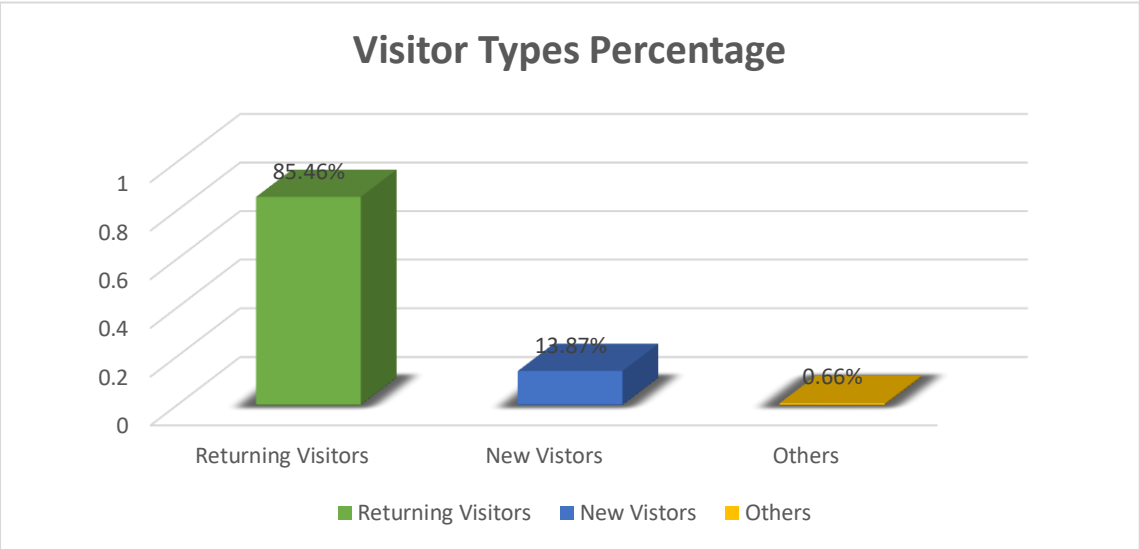




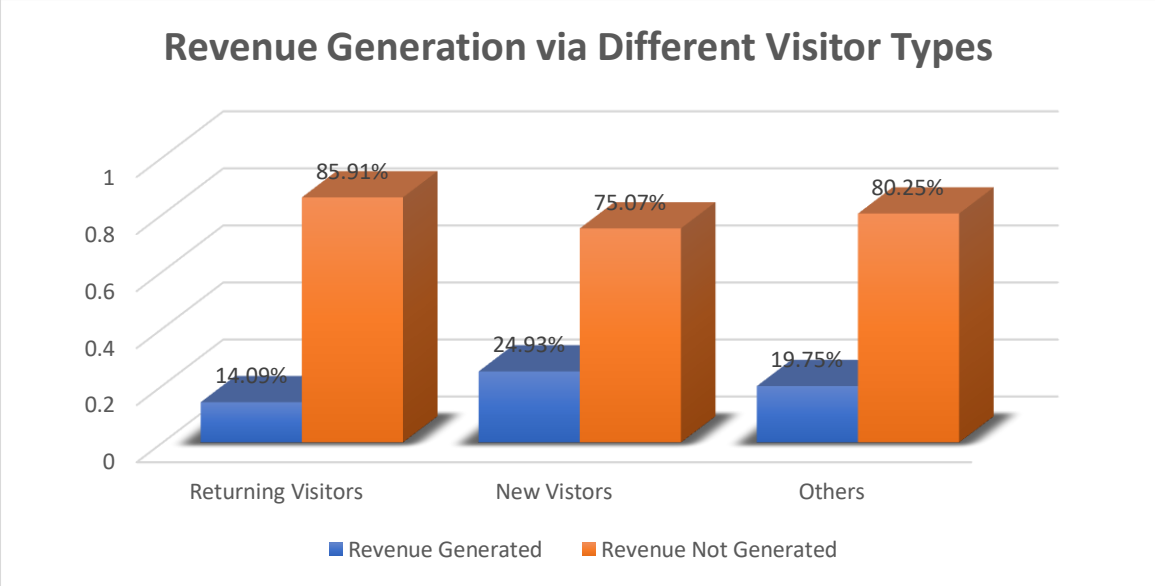
As we can visualize, the target column is ***highly imbalanced***. Most of the customers have not made any transaction. If we use this dataframe as the base for our predictive models and analysis we might get a lot of errors and our algorithms will probably overfit since it will "assume" that most of the customers will not make any transactions. We will treat this in the upcoming process.

Visitor Types vs Revenue

There are 3 different types of visitors: Returning Visitors, New Visitors and Others.



The number of new visitors are very less as compared to returning visitors.

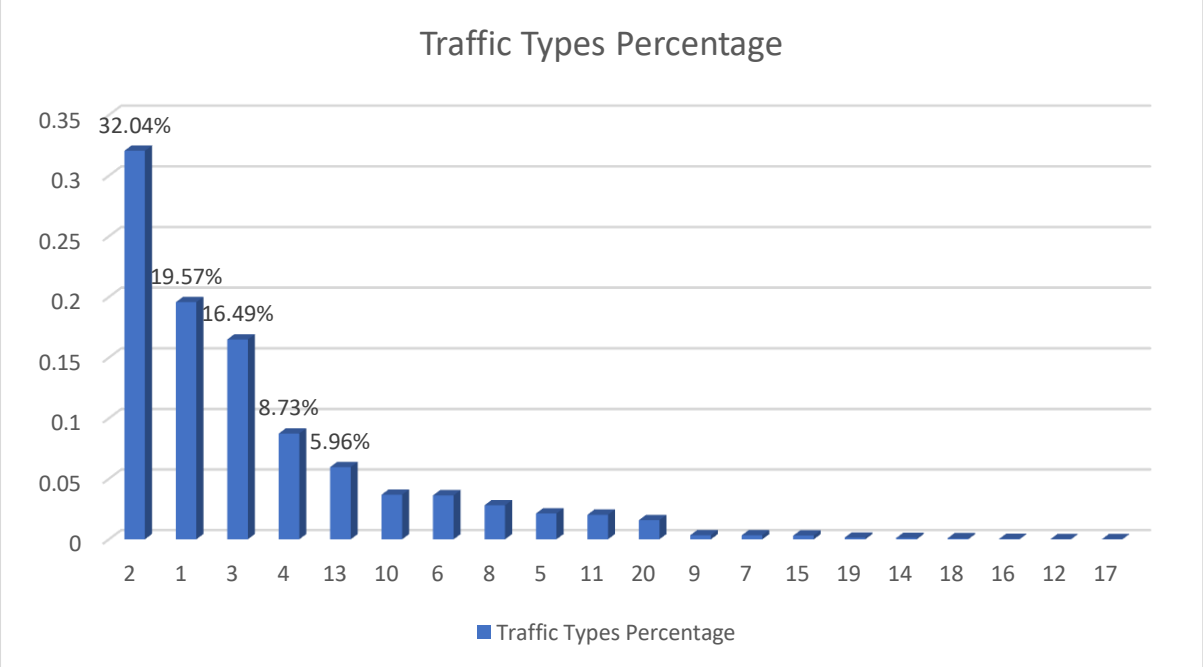


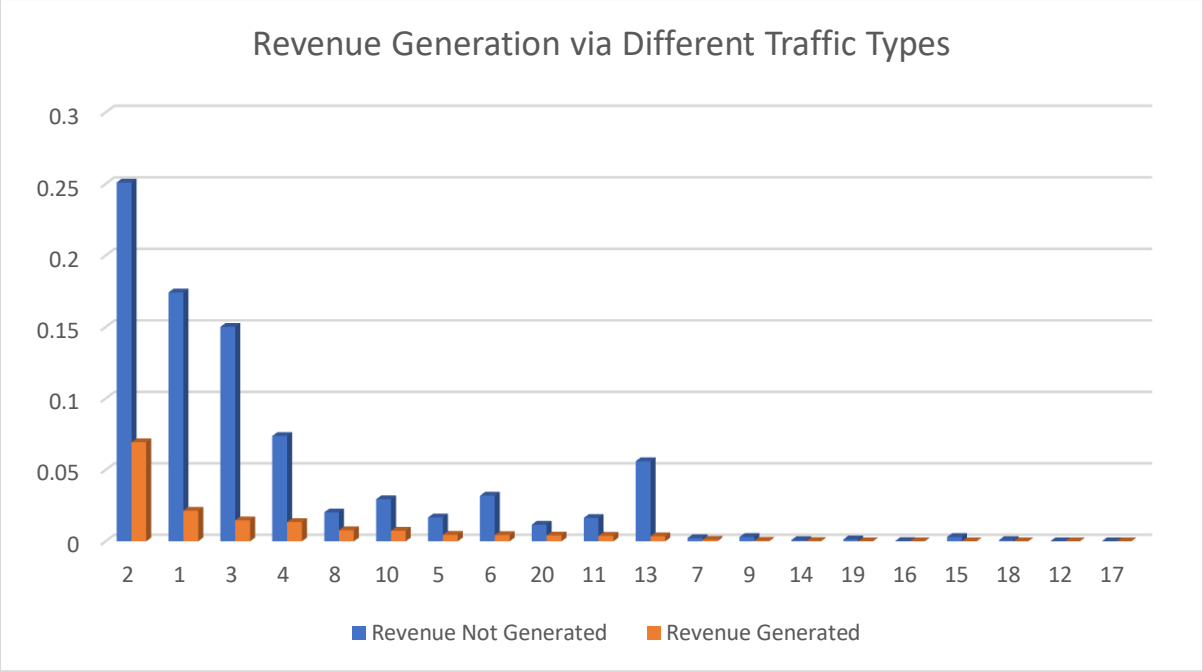
Visitor Type	Revenue Not Generated	Revenue Generated
New Visitor	75.07%	**24.93%**
Other	80.25%	19.75%
Returning Visitor	85.91%	**14.09%**

Though the returning visitors corresponds to ~85% of the overall sessions, but the conversion rate of new visitors is (24.93 – 14.09) ~ 10% greater than the returning visitors.

Traffic Types vs Revenue

There are 20 different traffic types ranging from 1 to 20.





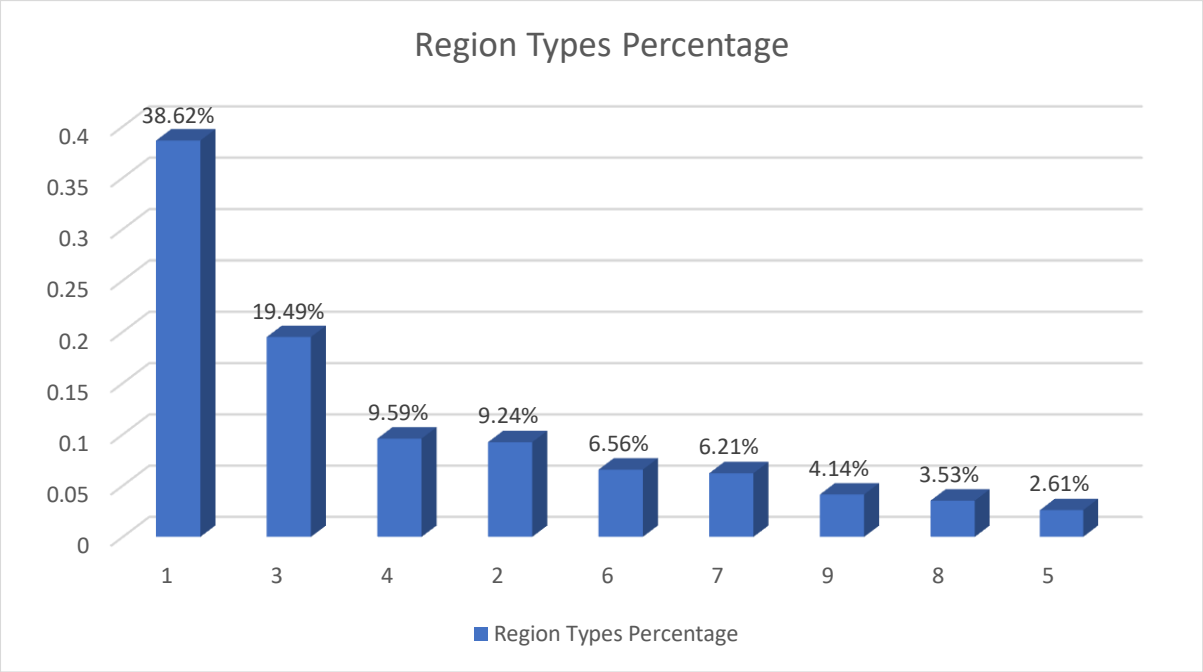
Traffic Types	Revenue Not Generated	Revenue Generated
1	89.03	10.97
2	78.34	21.66
3	91.06	8.94
4	84.52	15.48
5	78.46	21.54
6	88.04	11.96
7	70.00	30.00
8	72.30	27.70
9	90.24	9.76
10	80.00	20.00
11	80.97	19.03
12	100.00	0.00
13	94.09	5.91
14	84.62	15.38
15	100.00	0.00
16	66.67	33.33
17	100.00	0.00
18	100.00	0.00
19	94.12	5.88
20	74.09	25.91

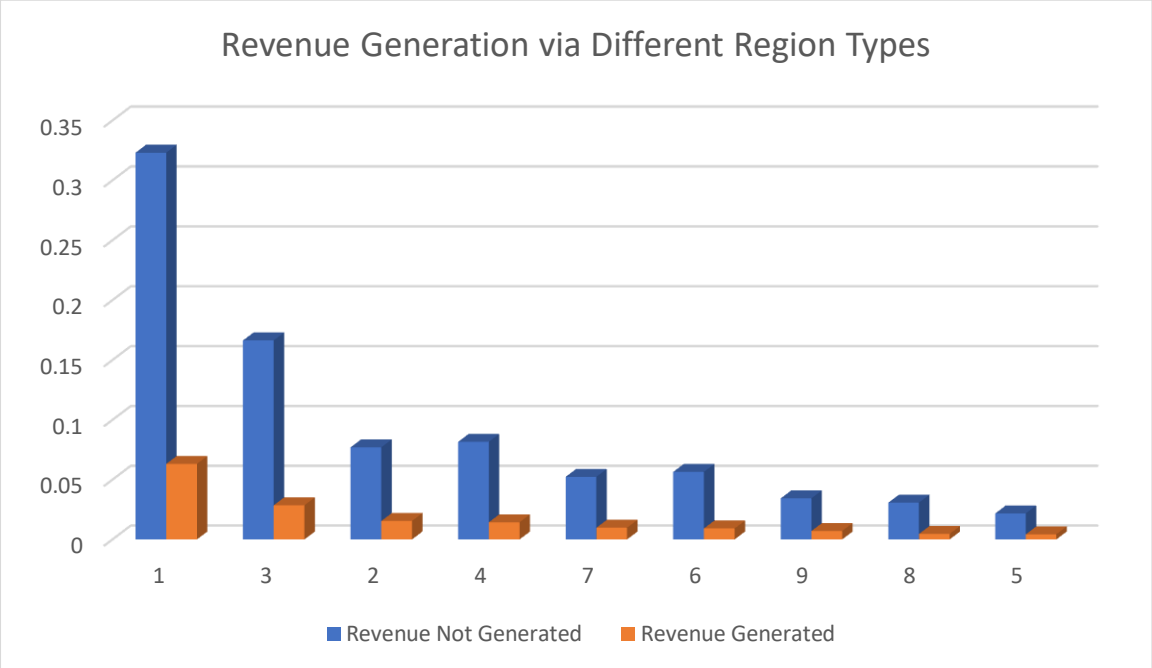
82.79% of the traffic generation is via Traffic Types 2, 1, 3, 4, and 13 respectively.
Overall, Traffic Type 2 has contributed more in the revenue generation w.r.t other traffic types.

Also, Traffic Types 12, 15, 17 and 18 has zero contribution in revenue generation. The company must look up what these traffic types corresponds to in real terms so that either they can improve the digital marketing if it corresponds to ads, social media etc. or ignore it if it is coming from something non-relevant.

Region vs Revenue

There are 9 different region types ranging from 1 to 9.

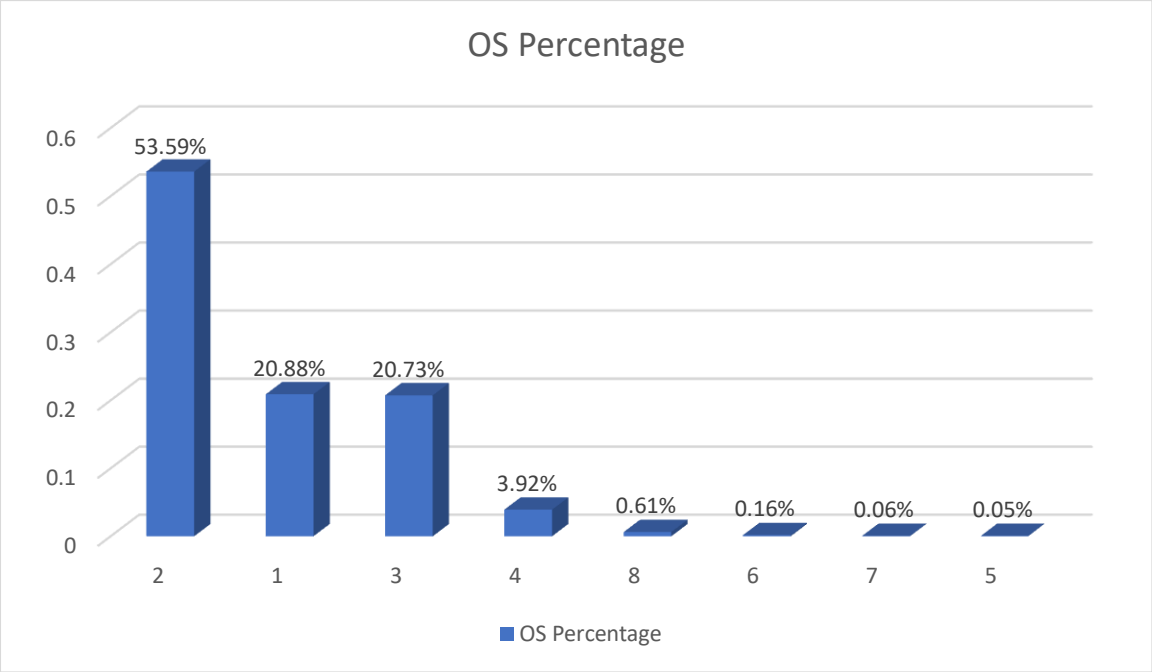


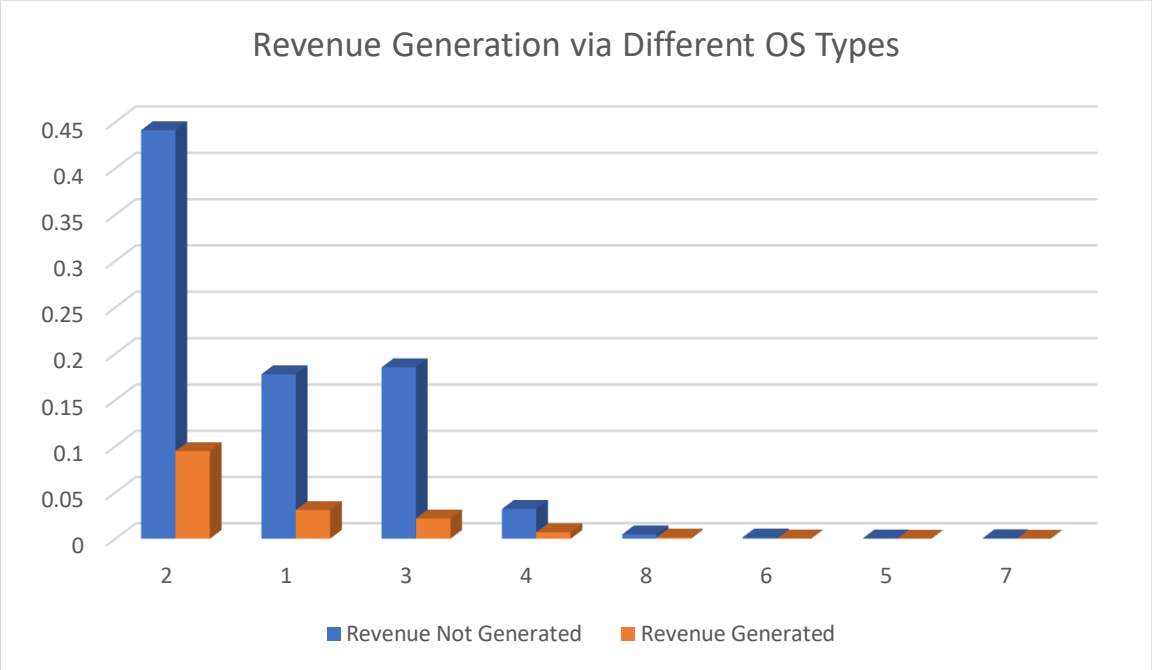


76.94% of the customers are from region 1, 3, 4, and 2 respectively and only customers from region 1 and 3 are contributing maximum in revenue generation. More efforts needs to be made in the region 2 and 4 to increase the revenue generation.

Operating System vs Revenue

There are 8 different operating system types ranging from 1 to 8.

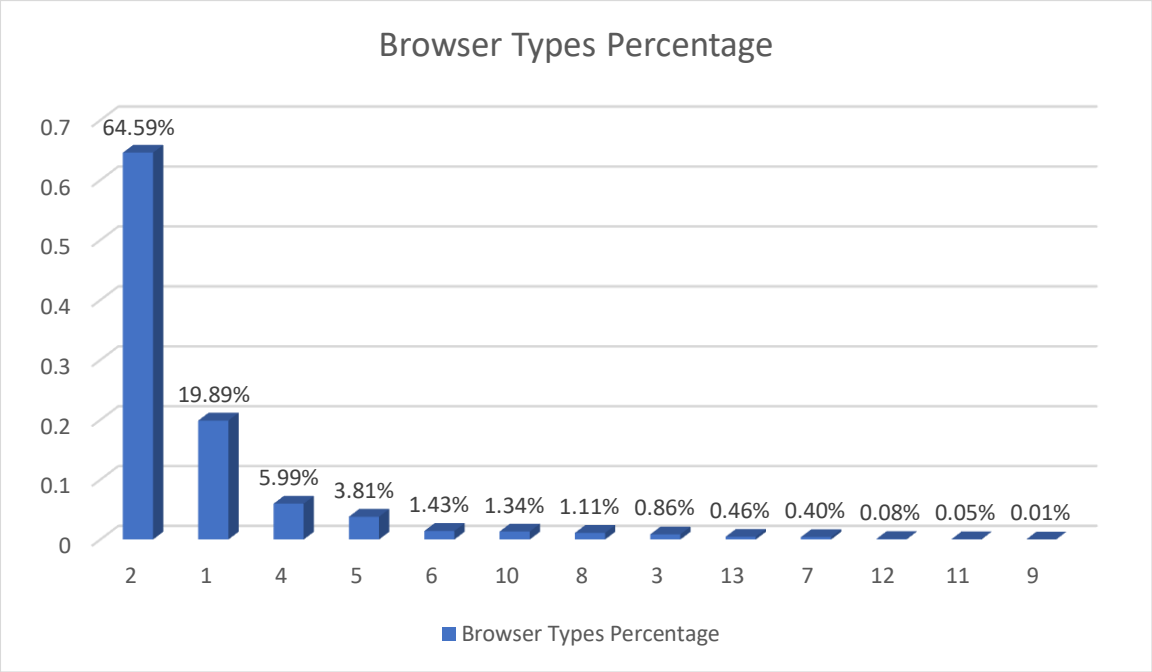


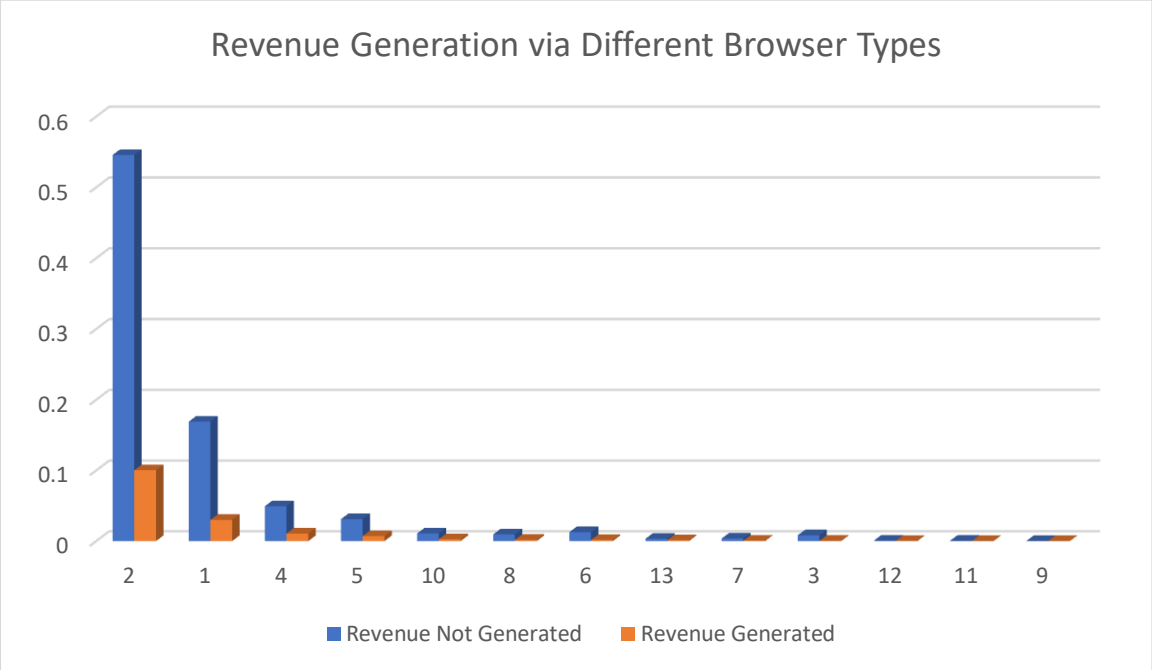


95.2% of the customers/visitors are having OS types 2, 1, and 3 respectively. Conversion rate is high on OS type 2 in comparison with type 1 and 3.

Browser vs Revenue

There are 13 different browser types ranging from 1 to 13.

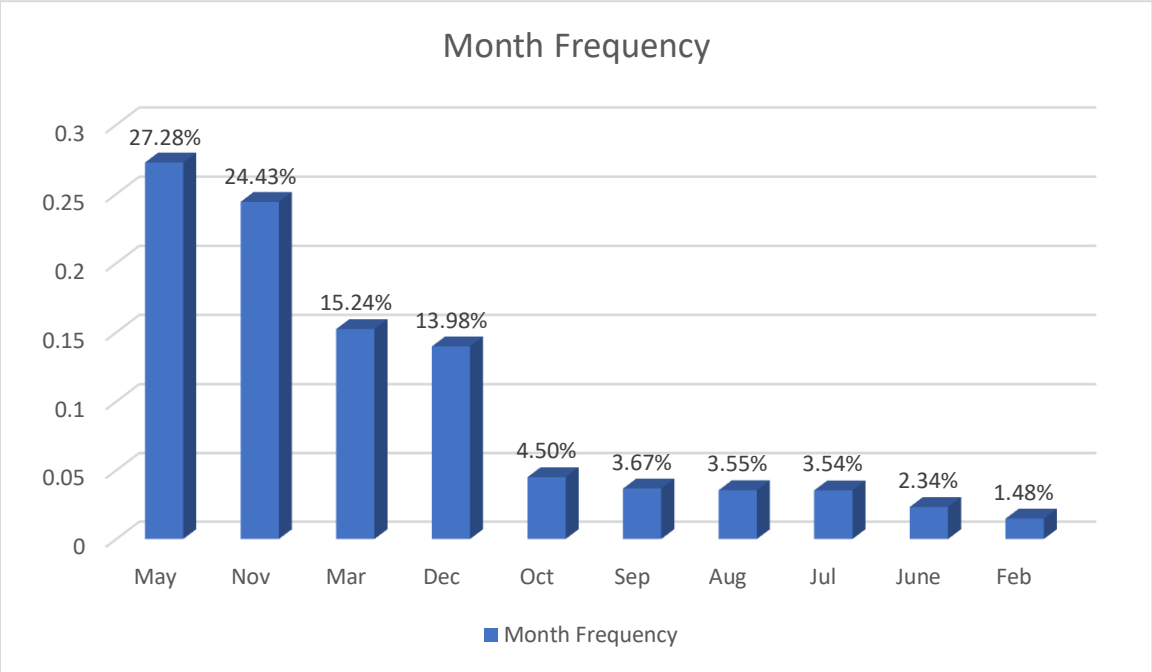


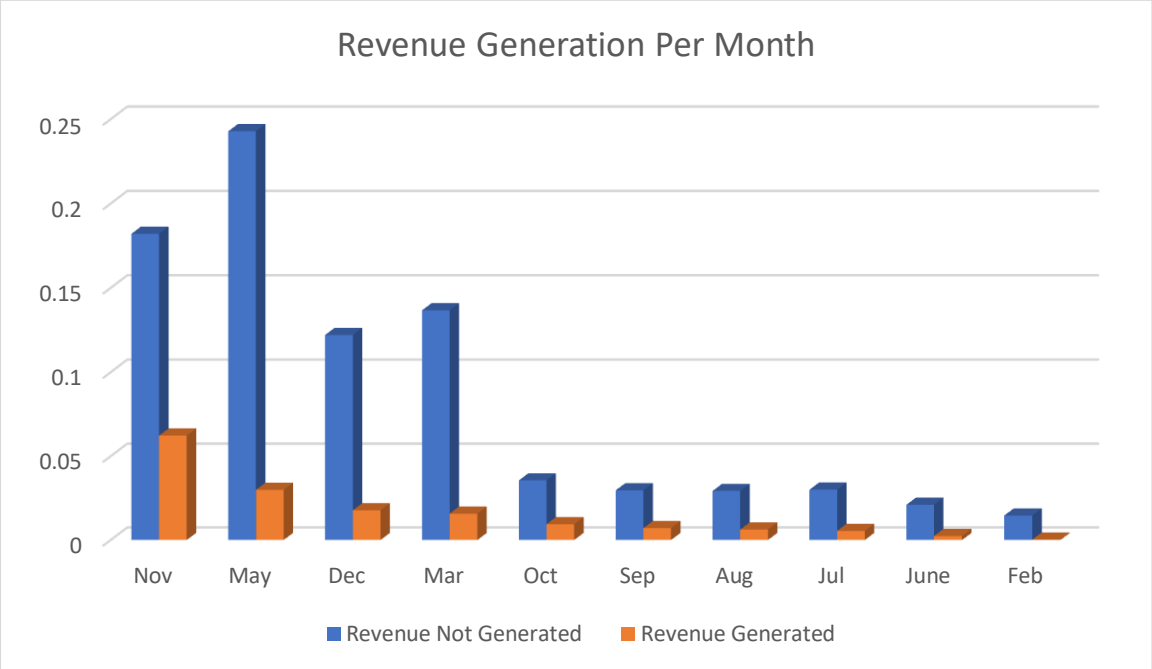


84.48% of the customers/visitors are using Browser type 2 and 1 respectively. Apart from browser type 4 and 5, all other browsers have very less usage rate.
Browser type 2 has a contribution of 10% in revenue generation overall.

Month vs Revenue

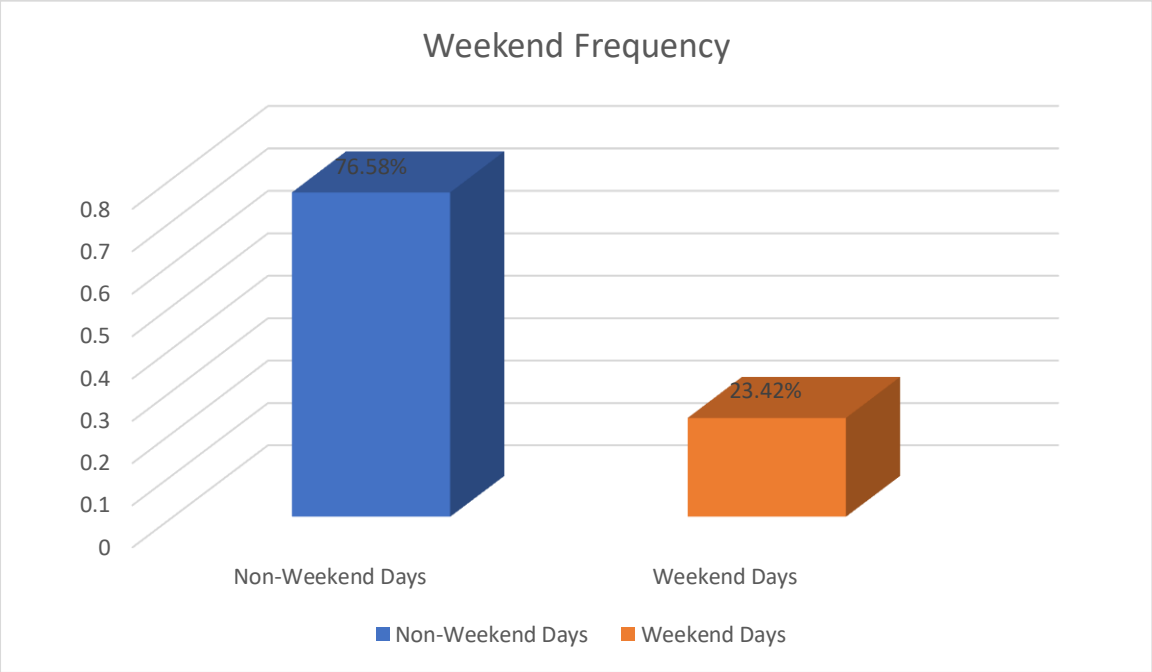
There are only 10 months of records in the given data. January and April has no records.

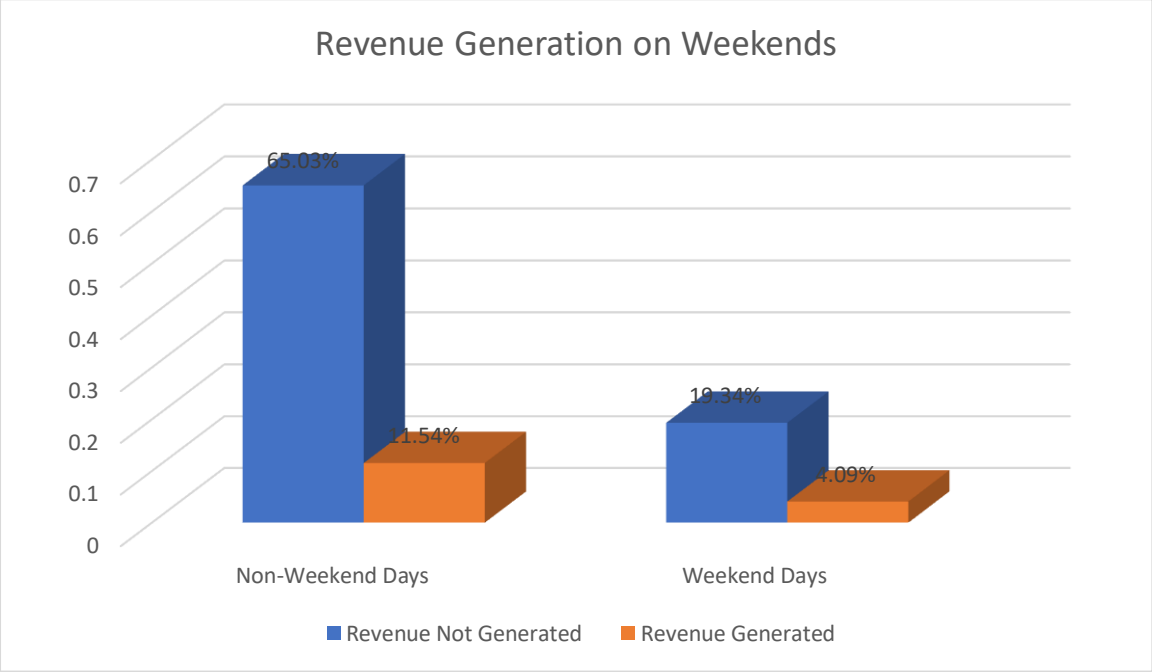




80.93% of customers were active in the months of May, November, March, and December combined.
November has contributed the maximum in terms of revenue generation while February has the least contribution in the revenue generation.
Although, May has the highest number of visitors but the revenue generation is approximately 52% more in November in comparison to May.

Weekend vs Revenue

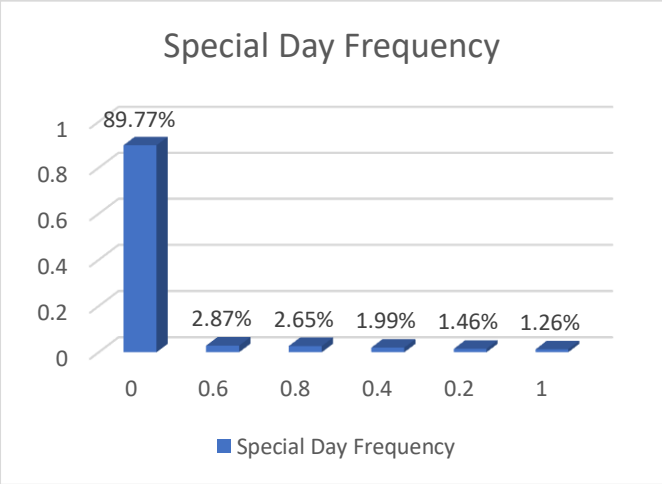


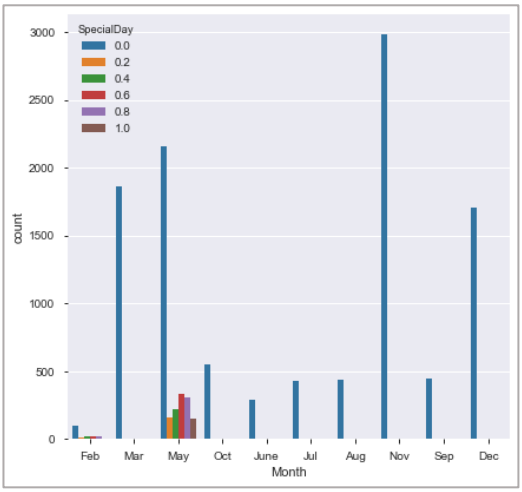


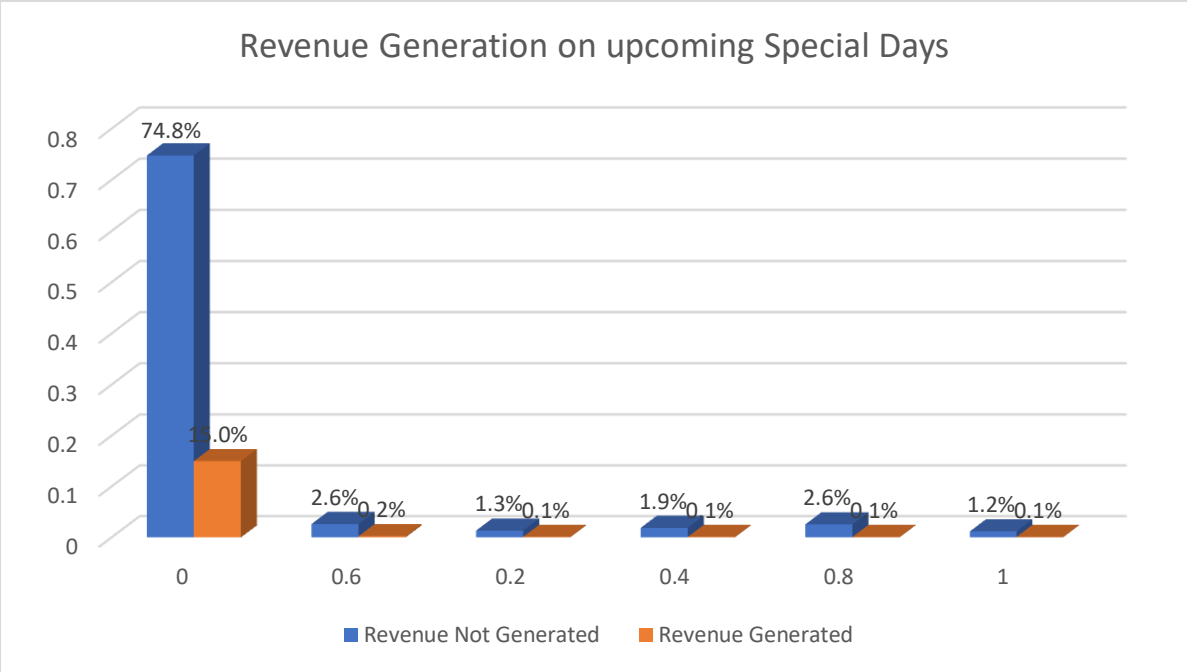
Weekday Type	Revenue Not Generated	Revenue Generated
Non-Weekend	84.92%	15.08%
Weekend	82.55%	17.45%

Based on the data, it seems like customers/visitors like to shop on non-weekends more in comparison to weekends.
But revenue conversion is slightly higher on weekends.

Special Day vs Revenue







Approximately 90% of the interactions or session generation happened on Non-Special Days. Since it's a Sportswear Company, there is no affinity for Special Days to Revenue generation.

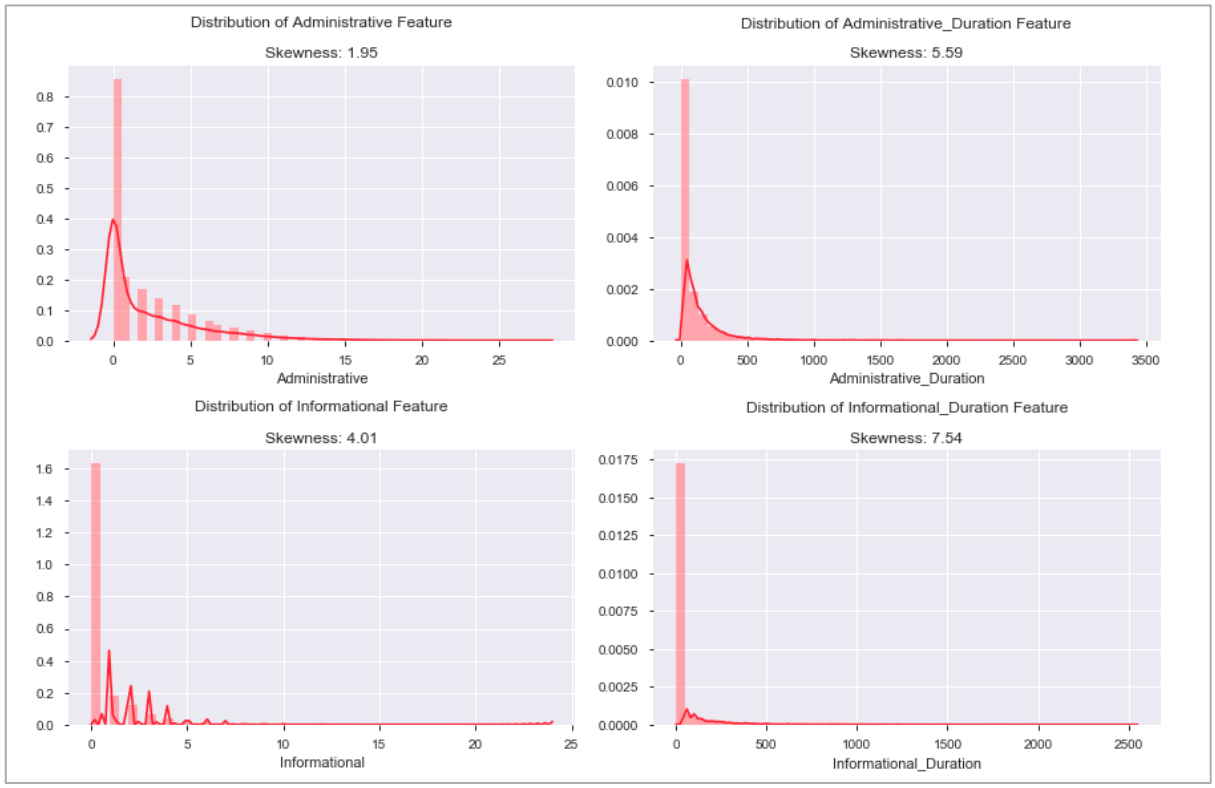
Also, most of the special days are in the month of May followed by February.

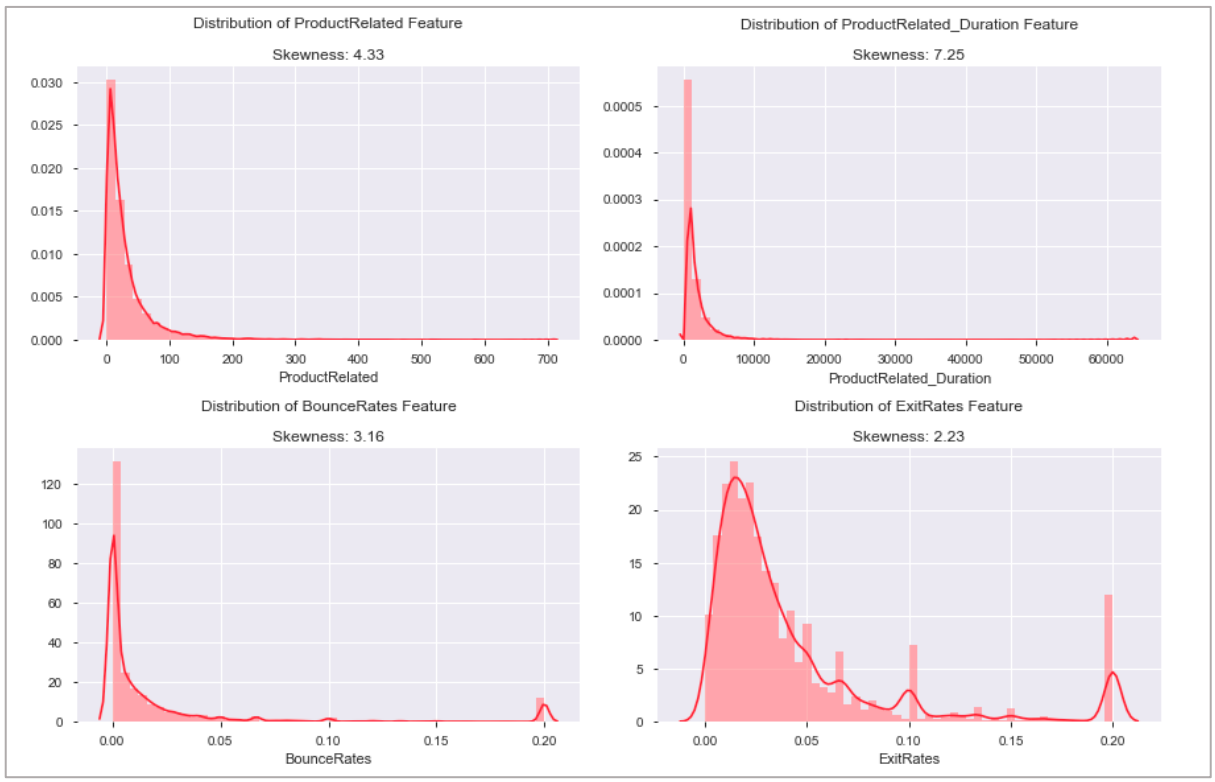
Administrative vs Informational vs ProductRelated

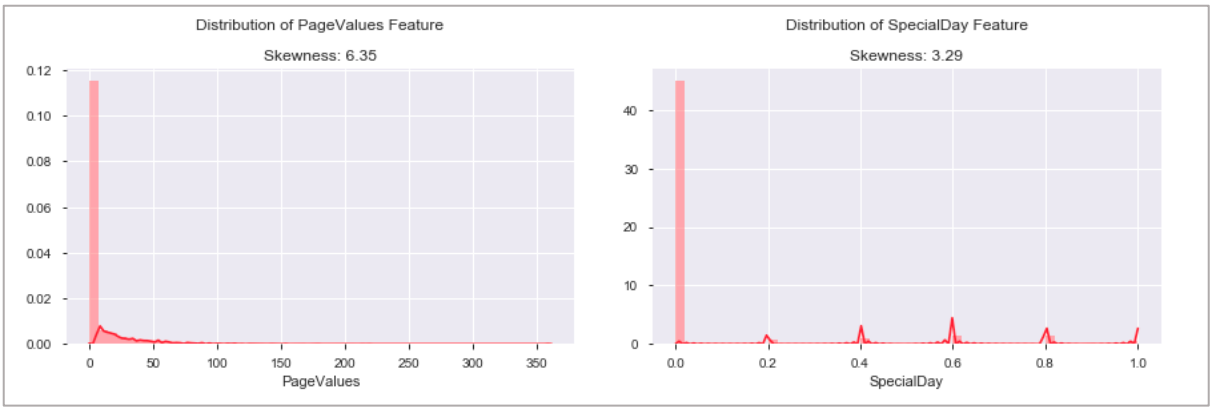
Page Types	Average Time Spent
Administrative	81.65
Informational	34.83
Product Related	1206.98

Average time spent on Product Related pages is more than Administrative and Informational Pages.

Below is the Distribution of the following Numerical Features along with their skewness: 'Administrative', 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay'.







All the numerical features are highly skewed. To measure which transformation works better, we created a table having the original skewness values of the numerical features and did a comparative analysis by applying log, square root, cube root, 4^{th} power, 5^{th}, 6^{th}, and 7^{th} power transformations and applied the one having the minimum skewness values.

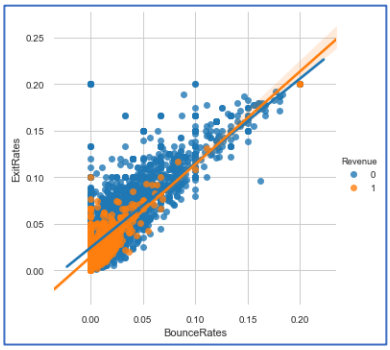
Features	Skewness	Log_T	Sqrt_T	Cbrt_T	4thPower_T	5thPower_T	6thPower_T	7thPower_T
Administrative	1.95	0.55	0.63	0.24	0.08	0	-0.04	-0.07
Administrative Duration	5.59	0.24	1.53	0.69	0.37	0.21	0.12	0.06
Informational	4.01	2.01	1.93	1.62	1.51	1.46	1.44	1.42
Informational Duration	7.54	1.9	3.42	2.43	2.05	1.87	1.77	1.7
ProductRelated	4.33	-0.04	1.5	0.81	0.45	0.19	-0.06	-0.31
ProductRelated Duration	7.25	-1.43	1.41	0.42	-0.21	-0.72	-1.17	-1.56
BounceRates	3.16	3.08	1.72	0.9	0.48	0.26	0.13	0.05
ExitRates	2.23	2.13	1.21	0.69	0.26	-0.17	-0.64	-1.15
PageValues	6.35	1.79	2.52	1.89	1.66	1.54	1.48	1.44
SpecialDay	3.29	3.1	2.86	2.74	2.7	2.67	2.66	2.65

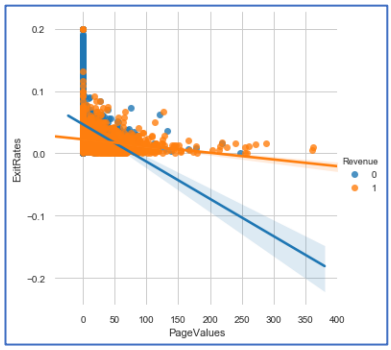
Since, the distribution of Special Day shows individual peaks at each interval, we will convert it to a categorical feature.

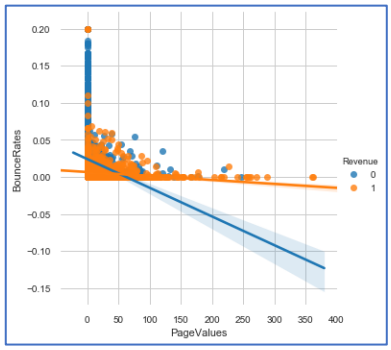
We applied sklearn's PowerTransformer to transform the data and StandardScaler and RobustScaler are used to scale the data.

We did comparison of multiple test cases and found that PowerTransformer with 'yeo-johnson' method and standardize is equal to False and with RobustScaler is giving a good score.

Bounce Rates vs ExitRates vs PageValues Correlation:

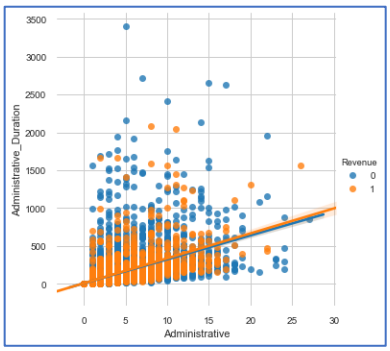


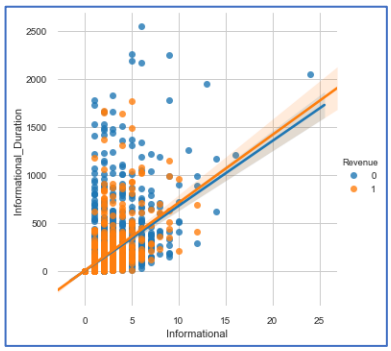


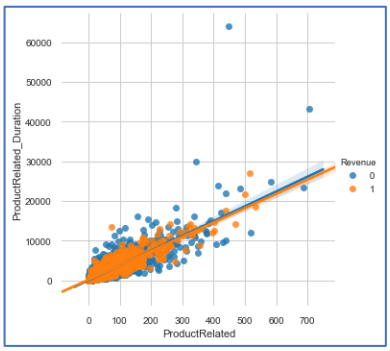


Exit Rates and Bounce Rates are positively correlated with the correlation coefficient of 0.90. Exit Rates and Page Values are negatively correlated. As the page value increases, exit rates decreases and vice-versa.
Bounce Rates and Page Values are negatively correlated. As the page value increases, exit rates decreases and vice-versa.

Administrative vs Administrative_Duration, Informational vs Informational_Duration vs ProductRelated vs ProductRelated_Duration





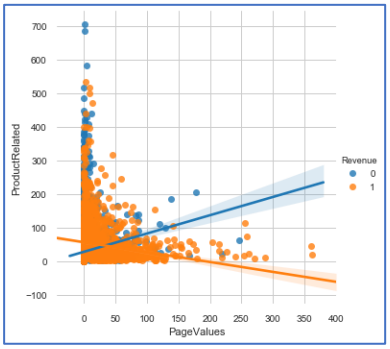


Administrative Pages and Administrative Duration, Informational and Informational Duration, Product Related and Product Related Duration are positively correlated.

On administrative pages 2 to 15 such as login, logout, password recovery, profile, email wish list etc., visitors have spent more than 500 seconds (approx. 8 minutes) which is generally quite higher than normal. It suggests that visitors are having trouble logging in or it's taking too much time to process the request.

Even though customers/visitors have spent a large amount of time on product related pages but the revenue generation is very low. There are certain outliers who did spent more than 60000 seconds (approx. 17 Hours) but still didn't make any transaction.

With increasing Page Values, the revenue generation is more. There are certain pages which have very less page values which need to be improved in order to generate revenue.



Based on the exploratory data analysis, we did a Q/A to know more insights about the data. Below are the Q/A as per the data:

Q. Which TrafficTypes have generated no Revenue?

A. Traffic Types: 12, 15, 17, & 18

Q. Which TrafficType(s) has generated maximum Revenue?

A. Traffic Type: 2

Q. Which TrafficTypes have generated high bounce rate and what is the overall Revenue generation from those TrafficTypes?

A. Traffic Types: 11 (2), 20 (1), and 8 (1)

Traffic types 11, 20, and 8 accounts for 24.52% of the overall revenue generation.

Bounce rate can be wildly different depending on the source of traffic. For example, it's likely that search traffic will produce a low bounce rate while social and display traffic might produce a high bounce rate.

Q. Which TrafficTypes generate high exit rate and what is the overall Revenue generation from those TrafficTypes?

A. Traffic Types: 11 (1), 20 (1), and 8 (1)

Traffic types 11, 20, and 8 accounts for 24.52% of the overall revenue generation.

Q. Which Region generates high bounce rate?

A. Region 1 and 3.

Q. What are the top 3 TrafficTypes of New visitors?

A. Traffic Types 2, 8 and 5.

Q. What are the maximum time spent on Administrative, Informational and ProductRelated pages and what are the page numbers?

A. 17.48 Hours were spent on the product related page number 449.

56.65 Minutes were spent on administrative page number 5.

42.49 Minutes were spent on informational page number 6.

Even though the revenue was not generated in the above case.

MODELLING TECHNIQUE

It's important to establish a naive baseline before we begin making machine learning models. If the models we build cannot outperform a naive guess then we might have to admit that machine learning is not suited for this problem. This could be because we are not using the right models, because we need more data, or because there is a simpler solution that does not require machine learning. Establishing a baseline is crucial so we do not end up building a machine learning model only to realize we can't actually solve the problem.

For a classification task, a good naive baseline is to run our model on simple cleaned data (i.e, original data with no null values), categorical features encoded and scaled. If after outlier treatment, EDA, feature engineering, feature selection our models cannot do better than baseline then we need to rethink our approach.

Metric: F1 Score :

There are a number of metrics used in machine learning tasks and it can be difficult to know which one to choose. Most of the time it will depend on the particular problem and if you have a specific goal to optimize for. Rather than calculating multiple metrics and trying to determine how important each one is, we should use a single number.

Since, our classification problem is to predict the intent of the visitor i.e., whether the visitor will make a transaction or not, we will be using ‘F1 score as our evaluation metric’.

Since our data is highly imbalanced and predict probability of our model would be rightly skewed and most of the prediction would be converging towards 0 (in other words – our model would be predicting that most of the visitors will abandon the site without making any transaction), which is why using accuracy won't give us the correct picture as our overall accuracy is going to increase because of the correct prediction of the True Negatives.

Base Model:

We performed the Logistic Regression classification on our dataset without making our data any more optimized for the machine to predict better classes. We will also not fine tune the hyper parameters for this algorithm as we are aiming to establish a baseline and get a sense of direction. After applying the Logistic Regression, we got the following classification report:

	precision	recall	f1-score	support
0	0.93	0.95	0.94	3116
1	0.68	0.58	0.62	546
accuracy			0.90	3662
macro avg	0.80	0.76	0.78	3662
weighted avg	0.89	0.90	0.89	3662

In this case, the F1-score is 62% which will act as our baseline for future models.

We also applied Random Forest Classifier and got the following classification report:

	precision	recall	f1-score	support
0	0.93	0.97	0.95	3116
1	0.75	0.57	0.65	546
accuracy			0.91	3662
macro avg	0.84	0.77	0.80	3662
weighted avg	0.90	0.91	0.90	3662

In this case, the F1-score increased a bit and is 65%. We have established a baseline metric so we can determine if our model is better than guessing!

Model Building:

Now, the dataset has been transformed using Power Transformer and scaled using Robust Scaler. The dataset is fed to Logistic Regression, Decision Tree, Random Forest, Gradient Boosting and Light Gradient Boosting classifiers using ten-fold cross validations. The Accuracy, Precision, Recall, and F1-Score are presented for each classifier.

Algorithms	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.89559	0.69468	0.61602	0.65233
Decision Tree	0.86281	0.5684	0.57422	0.56298
Random Forest	0.88997	0.73771	0.49926	0.61058
Gradient Boosting	0.89828	0.71294	0.60428	0.65325
Light Gradient Boosting	0.89863	0.71908	0.59695	0.65159

Results Obtained on Imbalanced Dataset:

To deal with class imbalance problem, we used oversampling method, in which a uniform distribution over the classes is aimed to be achieved by adding more of the minority (positive class in our dataset) class instances. Since this dataset is created by selecting multiple instances of the minority class more than once, first oversampling the dataset and then dividing it into training and test sets may lead to biased results due to the possibility that the same minority class instance may be used both for training and test. For this reason, in our study, 30 percentage of the data set consisting of 3662 samples out of 12205 are left out for testing and the oversampling method is applied to the remaining 70 percent of the samples.

Algorithms	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.89842	0.66926	0.63004	0.64906
Decision Tree	0.86264	0.53589	0.58791	0.56070
Random Forest	0.89869	0.72846	0.51099	0.60065
Gradient Boosting	0.90715	0.71915	0.61905	0.66535
Light Gradient Boosting	0.90688	0.72331	0.60806	0.66070

Results Obtained after oversampling:

Algorithms	F1 Score	Accuracy Score	Train Score	Test Score
Random Forest	0.90369	0.90770	1	0.90770
Gradient Boosting	0.90175	0.90470	0.93363	0.90470
Logistic Regression	0.89714	0.89842	0.89875	0.89842
Light Gradient Boosting	0.89420	0.89787	0.99684	0.89787
Decision Tree	0.86436	0.86100	1	0.86100