COSC2789 - Assignment 3

Topic 1.2: Online Shoppers Purchasing Intention

Group 23

Binh Nguyen - S3883631
Dung Nguyen - S3883630
Phuc Tran - S3911244

Jan 17, 2023

Contact

Mr. Binh - S3883631@rmit.edu.vn
Mr. Dung - S3883630@rmit.edu.vn
Mr. Phuc - S3911244@rmit.edu.vn

Table of Content

I. Abstract:

For this assignment, we will work on problem type 1, “Focusing on Data Modelling.” We use the “Online Shoppers Purchasing Intention Dataset Data Set” (Sakar et al., 2018), and we will model the data by treating it as a Classification and Clustering. We will use Logistic Regression models and Support Vector Machines (SVM) models for the Classification approach. For the Clustering approach, we will use K-means and DBSCAN models.

II. Introduction:

A. Data Set Information:

The dataset contained 12,330 sessions. The data include session page type access/duration (administrative, informational, etc.), the session's pageview insights (bounce/exit rates, time, etc.), the session's information (operating systems, browser, etc.), and the session’s purchase (revenue). Among these 12,330 sessions, 84.5% (10,422) were negative class samples that did not end with shopping, and the remainder (1908) were positive class samples that did end with shopping. Thus, to avoid any tendency to a specific campaign, special day, user profile, or period, the dataset was also structured so that each session belonged to a different user for a year (Sakar et al., 2018).

B. Attribute Information:

#	Column	Non-Null Count	Dtype
0	Administrative	12330 non-null	int64
1	Administrative_Duration	12330 non-null	float64
2	Informational	12330 non-null	int64
3	Informational_Duration	12330 non-null	float64
4	ProductRelated	12330 non-null	int64
5	ProductRelated_Duration	12330 non-null	float64
6	BounceRates	12330 non-null	float64
7	ExitRates	12330 non-null	float64
8	PageValues	12330 non-null	float64
9	SpecialDay	12330 non-null	float64
10	Month	12330 non-null	object
11	OperatingSystems	12330 non-null	int64
12	Browser	12330 non-null	int64
13	Region	12330 non-null	int64
14	TrafficType	12330 non-null	int64
15	VisitorType	12330 non-null	object
16	Weekend	12330 non-null	bool
17	Revenue	12330 non-null	bool

Each session in the dataset contained 18 columns representing the following meanings:

- Administrative: page type administrative that the user visited.
- Administrative_Duration: the amount of time spent on administrative pages.
- Informational: page type informational that the user visited.
- Informational_Duration: the amount of time spent on informational pages.
- ProductRelated: page type product related that the user visited.
- ProductRelated_Duration: the amount of time spent on this product-related pages.
- BounceRates: percentage of visitors who enter the website without doing any additional tasks.
- ExitRates: percentage of pageviews on the website that ends at that specific page.
- PageValues: average page value averaged over the target page's value.
- SpecialDay: value the closeness of the browsing data to particular days or holidays.
- Month: month of the pageview occurred.
- OperatingSystems: operating system used to view the page.
- Browser: browser used to view the page.
- Region: the region in which the user is located.
- TrafficType: Categorized users' type of traffic.

- VisitorType: representing whether a visitor is New Visitor, Returning Visitor, or Other.
- Weekend: representing whether the session is on the weekend.
- Revenue: representing whether or not the user completed the purchase.

C. The goal of the project:

With the dataset of customer insights when accessing the e-com website, we can analyze the customer's consumption behavior through information from each session. From the information of each website visit session, we will be able to analyze more deeply to understand customer insights and what makes them decide to buy. For each attribute of the sessions, we can analyze the effectiveness in enticing customers to buy. From there, we will draw the most general views to plan marketing campaigns more effectively and bring the best outcome.

In addition, understanding the customer insights that visit the website also helps us reduce the website's target audience. From there, we can have promotional plans focusing on the target audience we analyzed through the existing dataset to drive better website traffic and increase purchase rates.

II. Methodology:

A. Exploratory Data Analysis:

1. Data Description:

	count	mean	std	min	25%	50%	75%	max
Administrative	12205.0	2.338878	3.330436	0.0	0.000000	1.000000	4.000000	27.000000
Administrative_Duration	12205.0	81.646331	177.491845	0.0	0.000000	9.000000	94.700000	3398.750000
Informational	12205.0	0.508726	1.275617	0.0	0.000000	0.000000	0.000000	24.000000
ProductRelated	12205.0	32.045637	44.593649	0.0	8.000000	18.000000	38.000000	705.000000
ProductRelated_Duration	12205.0	1206.982457	1919.601400	0.0	193.000000	608.942857	1477.154762	63973.522230
BounceRates	12205.0	0.020370	0.045255	0.0	0.000000	0.002899	0.016667	0.200000
ExitRates	12205.0	0.041466	0.046163	0.0	0.014231	0.025000	0.048529	0.200000
PageValues	12205.0	5.949574	18.653671	0.0	0.000000	0.000000	0.000000	361.763742
SpecialDay	12205.0	0.061942	0.199666	0.0	0.000000	0.000000	0.000000	1.000000

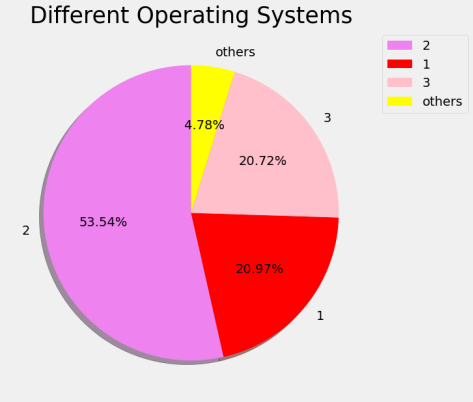
Administrative	numerical, quantitative, ratio	Min: 0, Max: 27
Administrative_Duration		Min: 0, Max: 3398.75
Informational		Min: 0, Max: 24
Informational_Duration		Min: 0, Max: 2549.3750
ProductRelated		Min: 0, Max: 705
ProductRelated_Duration		Min: 0, Max: 63973.52223
BounceRates		Min: 0, Max: 0.2
ExitRates		Min: 0, Max: 0.2
PageValues		Min: 0, Max: 0.2
SpecialDay	categorical, qualitative, nominal	No/Yes: 0/1
Month	string of months	

OperatingSystems	categorical, qualitative, interval	Min: 1, Max: 8
Browser		Min: 1, Max: 13
Region		Min: 1, Max: 9
TrafficType		Min: 1, Max: 20
VisitorType	string representing the type of visitors (New/Returning Visitor, Other)	
Weekend	boolean	True/False
Revenue	boolean	True/False

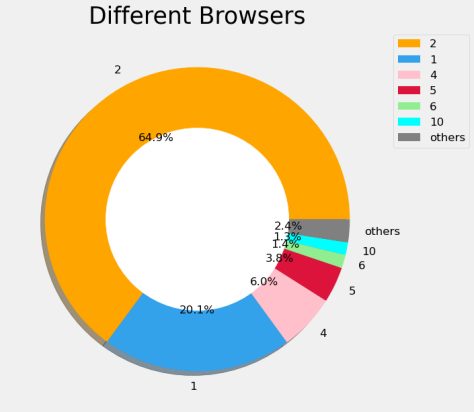
2. Graph and Analyze:

With this dataset, we will analyze and graph the essential attributes to get a better overview of the analyzed data. The primary attributes of sessions will include Operating Systems, Browsers, Time, and Visitor Type. In addition, we will also analyze Customer distribution data to understand better customers accessing the website. From the graphs of the primary attributes, we can combine the data with other attributes that are more in-depth and have a more detailed analysis of the revenue outcome.

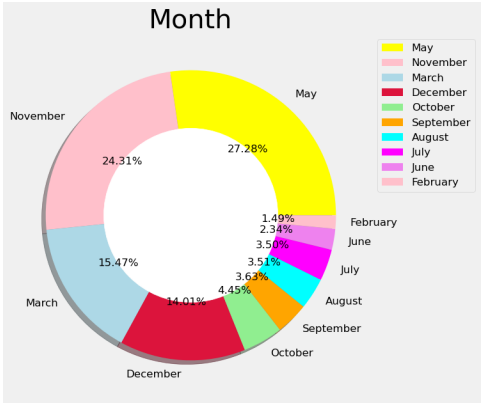
a) Sessions analysis:



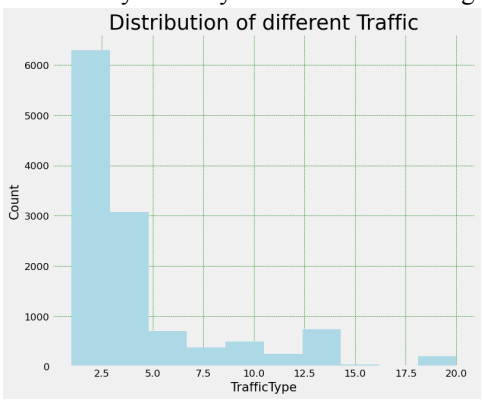
The top 3 Operating Systems take up approximately 95% of this data. Therefore we can know the leading sessions' OS we can focus on analyzing.



Of 12,330 sessions, 90% use these top 3 browsers to access the website: 64.9% for browser 2, 20.1% for browser 1, and 6.0% for browser 4. From these percentages, we can know what the favorite browsers customers use are.

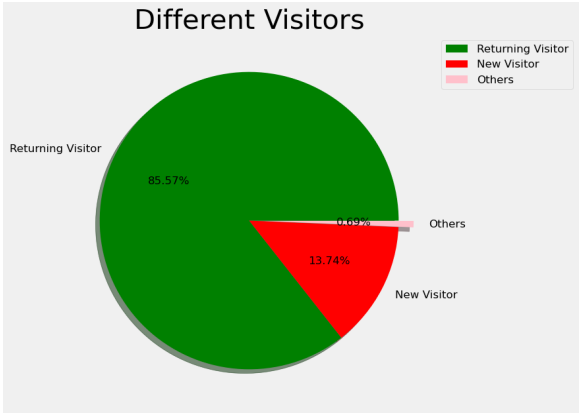


The website recorded the highest traffic in May, November, March, and December. We can analyze customers' favorite shopping period based on the time of the year many sessions are accessing the website.

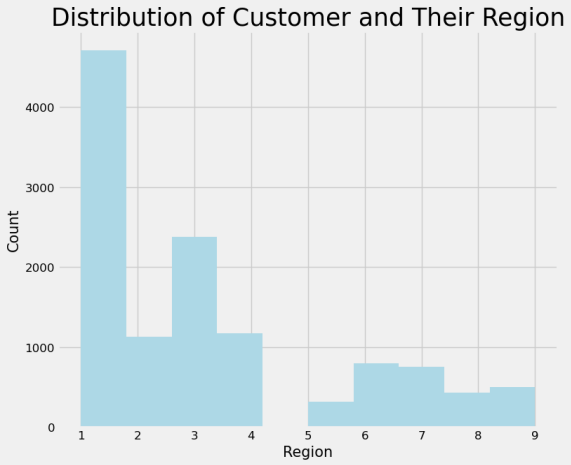


With 20 different Traffic Type distributions, we can see that the traffic is exponentially distributed. Since this data will help us better understand website traffic, we will use it to analyze further.

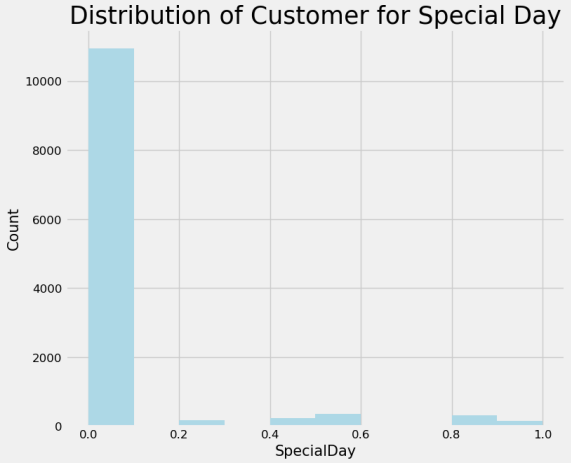
b) Visitors/Customers analysis:



More than 85% of sessions come from returning visitors. This number shows that this website has a high customer retention rate and a certain number of loyal customers.

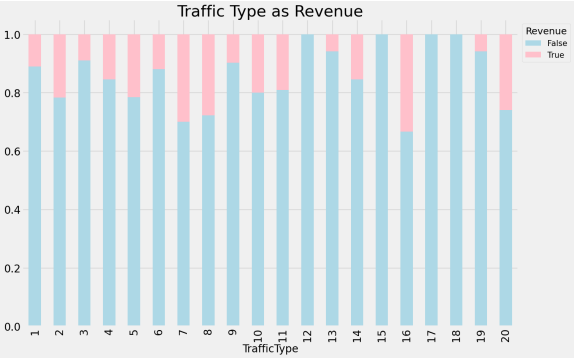


The customers who access the website are coming from 9 different regions. With the region distribution, we can see which region has the most traffic to the website and find the most favorable region to define the target audience for the marketing plan.



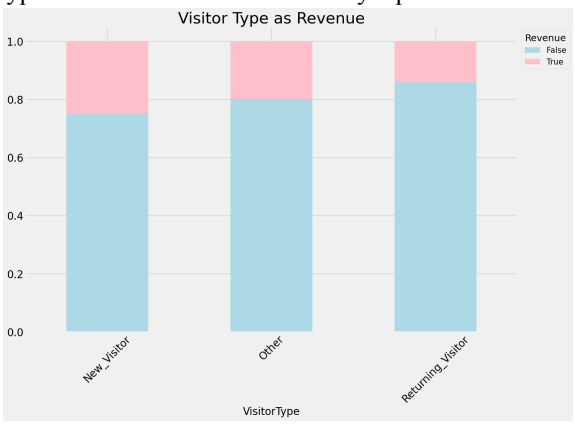
These are the sessions to access the website on a special day of the year. From the distribution of sessions on special days, we can analyze what marketing campaign is suitable for each particular day to attract customers.

c) Bivariate analysis:

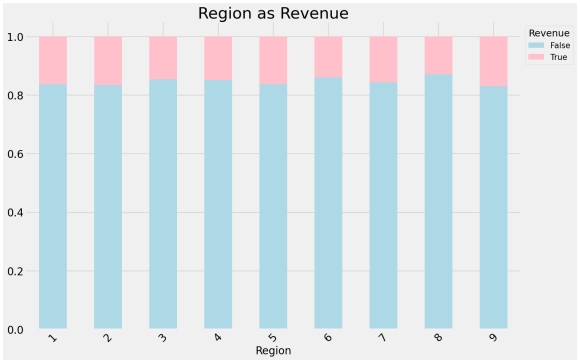


For this graph, we will treat TrafficType as categorical to analyze its revenue rate. Every category is different from others. Traffic types 7, 8, 16, and 20 have

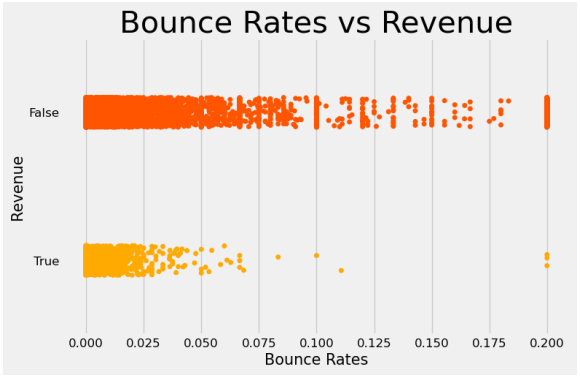
a higher rate than others and indicate that customers with those sessions' traffic types are easier to influence to buy a product.



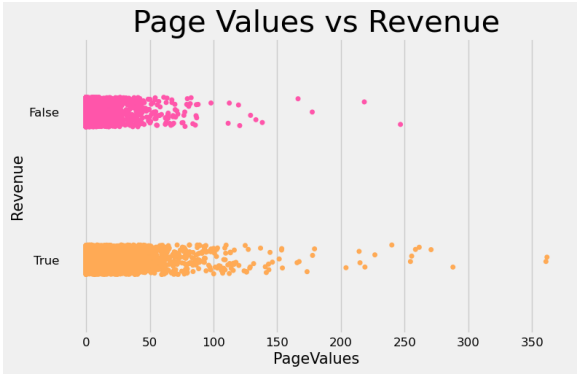
Here, we will analyze customer purchase rates based on VisitorType. As we can see, new visitors tend to purchase more than the rest of the visitor type.



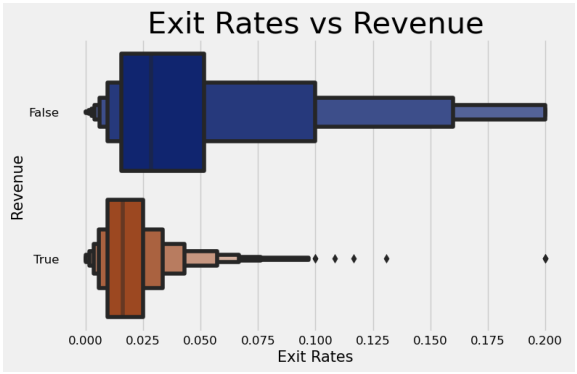
With this chart, we can analyze customers' purchasing trends according to the region they visit the website. All regions have relatively equal purchasing power.



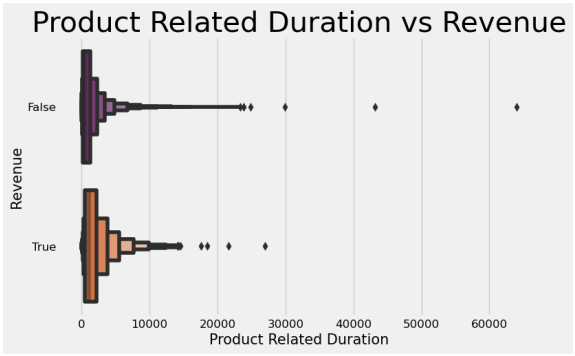
This chart shows that BounceRates is being distributed exponentially for two revenue trends on the website. There are also outliers showing non-purchase based on ExitRates. Moreover, BounceRate also has a significant influence on analyzing whether customers buy or not.

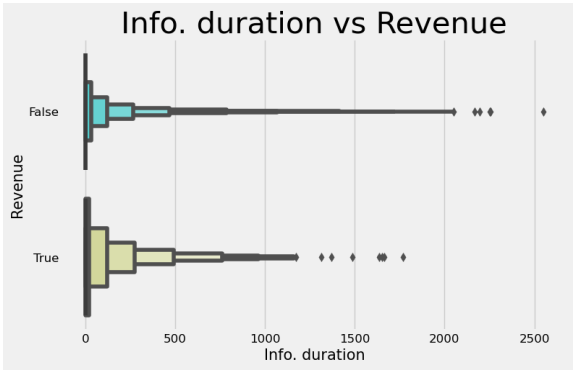


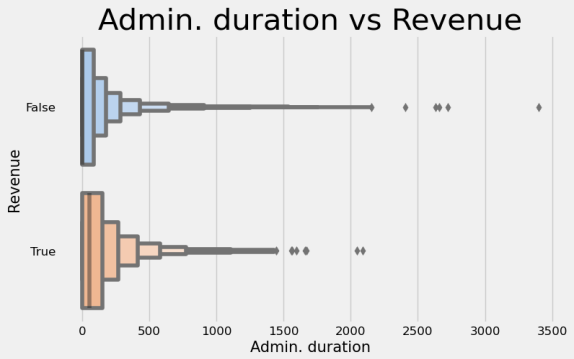
The distributed trend of PageValues is similar to the BounceRates in that it is exponential for both trends in revenue. However, in contrast to BounceRate, this chart shows many outliers in the yes-to-purchase revenue section. Therefore, PageValues will significantly influence the customers' buying rate on the website.



The ExitRates are typically (gaussian) distributed for both purchased(True) and not purchased (False). There are many outliers in not purchased (False).



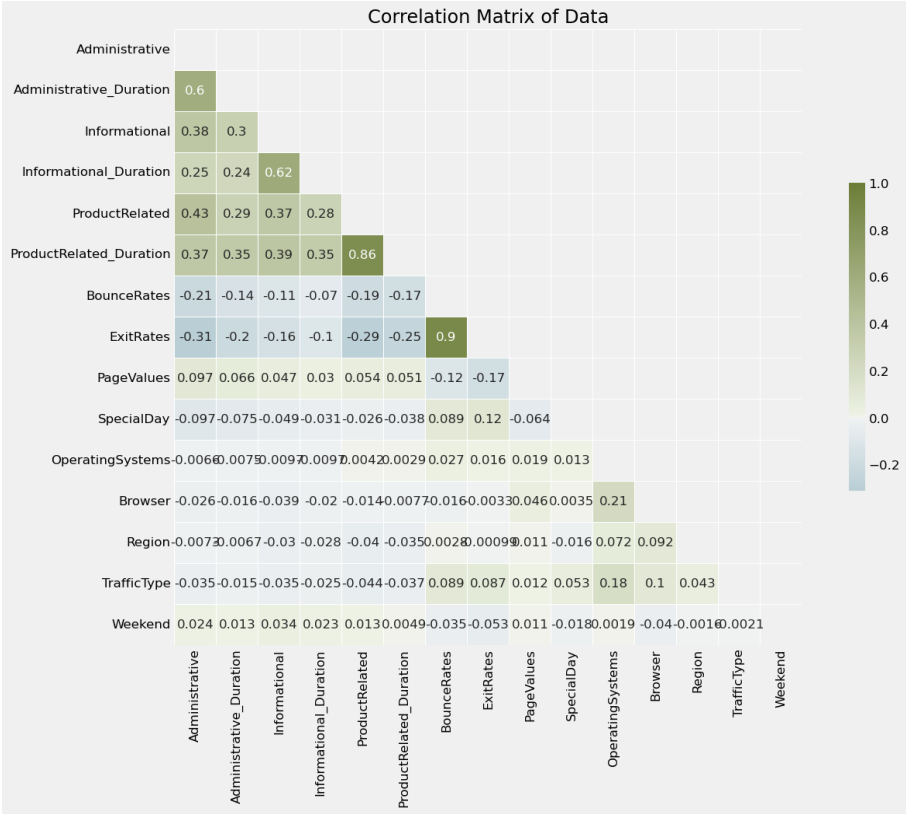




All three graphs show the correlation of revenue and ProductRelatedDuration, Informational_Duration, and Administrative_Duration are all exponentially distributed for both purchased (True) and not purchased (False). In addition, these charts also contain outliers in the non-purchase trend of revenue.

B. Modeling:

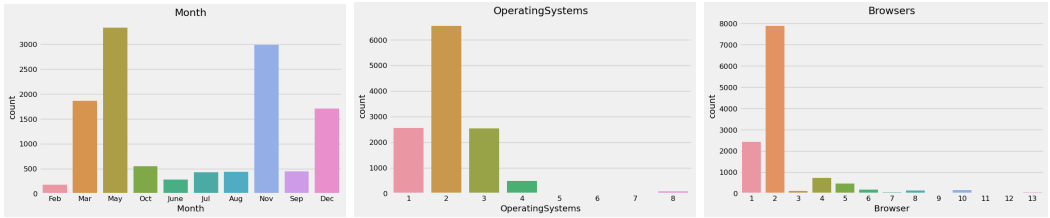
For further analysis, we will create the feature extractor to build a Machine Learning model. To build a good model, it is necessary to examine the raw input data. Technique correlation analysis will be performed and will help us inspect which column affects the most and the least to the final answer (which is the revenue).



- ProductRelated_Duration and ProductRelated are highly correlated with each other. This correlated trend can also be seen in ExitRates and BounceRates.
- The moderately correlated trend appears in these data pairs: Administrative_Duration/Administrative and Informational_Duration/Informational.

	Variable	Correlation
8	PageValues	0.491894
7	ExitRates	0.20432
4	ProductRelated	0.156042
5	ProductRelated_Duration	0.150077
6	BounceRates	0.145091
0	Administrative	0.13633
2	Informational	0.093626
1	Administrative_Duration	0.091768
9	SpecialDay	0.083601
3	Informational_Duration	0.069358
14	Weekend	0.027729
11	Browser	0.024052
10	OperatingSystems	0.014927
12	Region	0.012725
13	TrafficType	0.005618

If a variable has a very low correlation with the target variable (Revenue), it will not be helpful for the model prediction. Therefore, we will drop 'TrafficType,' 'Region,' 'OperatingSystems,' 'Browser,' 'Weekend,' and 'Informational_Duration' since these variables have a significantly small correlation with the Revenue.



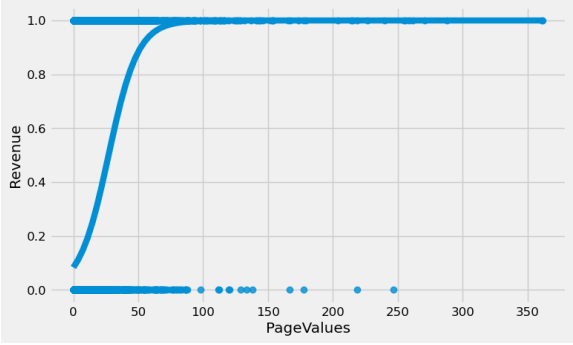
We can see that the 'Month' variable has only ten unique inputs, lacking two other months. Customers use only five operating systems, and two take most of the dataset. Similar to the 'Browser' column. So these variables will not be helpful in the model prediction and will drop to reduce the unnecessary features.

III. Results:

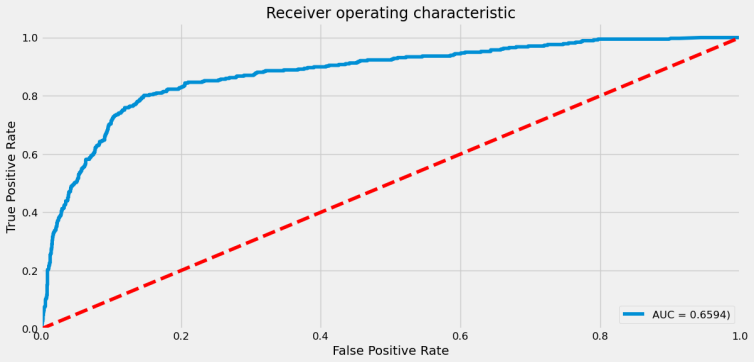
A. Classification:

We will use Two classification models: Logistic and SVM, to predict the likelihood of customers' behavior. We will split the data into training and test sets in the ratio of 80:20, respectively.

1. Logistic Regression:



Since the data is distributed logistically in these two columns, we will train the model based on logistic regression.




```
Training Accuracy:  0.8797623924621057
Testing Accuracy:  0.8807865628840639

              precision    recall  f1-score   support

       False       0.89      0.98      0.93      2063
        True       0.76      0.34      0.47       378

    accuracy                           0.88      2441
   macro avg       0.82      0.66      0.70      2441
weighted avg       0.87      0.88      0.86      2441


Confusion Matrix:
 [[2022   41]
 [ 250  128]]

Accuracy Score:  0.881
```

The Logistic model has a training accuracy of 87.9% and a testing accuracy of 88.0%.

2. Support Vector Machines (SVM):

```
Training Accuracy:  0.8461696026218762
Testing Accuracy:  0.8463744367062679
              precision    recall  f1-score   support

       False       0.85      1.00      0.92      2063
        True       1.00      0.01      0.02       378

    accuracy                           0.85      2441
   macro avg       0.92      0.50      0.47      2441
weighted avg       0.87      0.85      0.78      2441


Confusion Matrix:
 [[2063    0]
 [ 375    3]]

Accuracy Score:  0.846
```

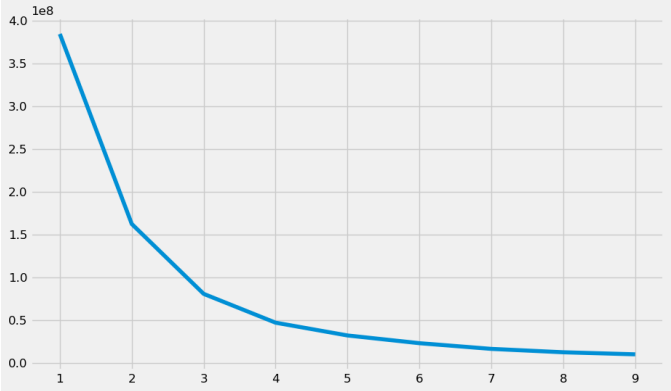
The SVM model has a training accuracy of 84.6% and a testing accuracy of 84.6%.

B. Clustering:

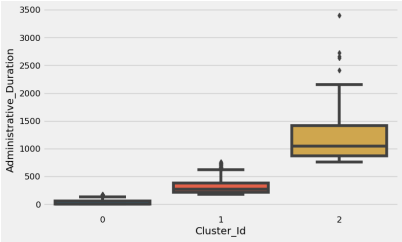
For clustering, we will use K-Means and DBSCAN models to predict the likelihood of customers' behavior. Thus, to have a better analysis, we will use three variables: Administrative_Duration, BounceRates, and ExitRates.

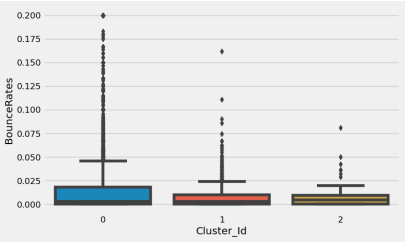
1. K-Means:

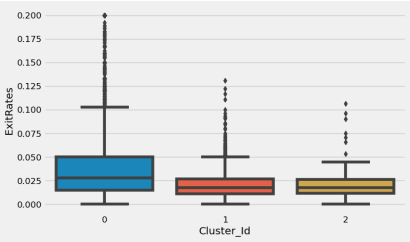
We use the Elbow curve method to determine the value to find the optimal number of clusters.



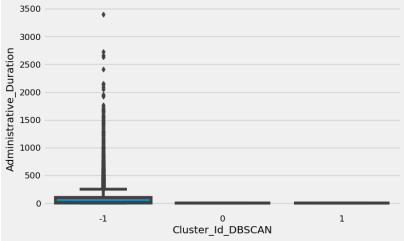
As we can see on the graph, the optimal number of clusters using the Elbow curve technique is 3. We can create a box plot from the three clusters to visualize Cluster Id vs. the three variables we mentioned before.

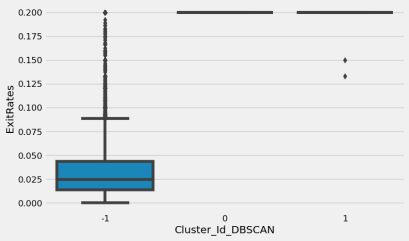


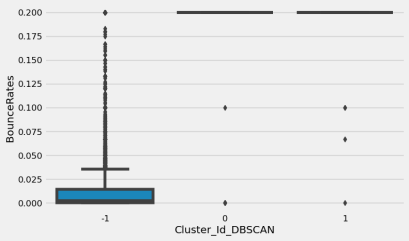




2. DBSCAN:







By using DBSCAN, we will better understand outliers' behavior in our dataset.

IV. Discussion:

A. Classification:

Most classification models work best when the number of classes is equal since they are designed to maximize accuracy and reduce error. Thus, they do not consider the class distribution/proportion or balance of classes. In our dataset, the percentage of customers purchasing on the website (class 1) is 15.5%, whereas 84.5% of customers did not buy anything (class 0).

We apply the confusion matrix to measure a classification algorithm's performance, containing information about the actual and predicted classes. The confusion matrix will calculate based on these metrics:

- Precision: How often is it correct when it predicts a positive result? i.e., limit the number of false positives.
- Recall: How often does it correctly predict when it is the positive result? i.e., limit the number of false negatives.
- f1-score: Harmonic mean of precision and recall.

The confusion matrix for class 1 (Purchased) would look like this:

	Predicted: 0 (Not Purchased)	**Predicted: 1 (Purchased)**
Actual: 0 (Not Purchased)	True Negatives	False Positives
Predicted: 1 (Purchased)	False Positives	True Negatives

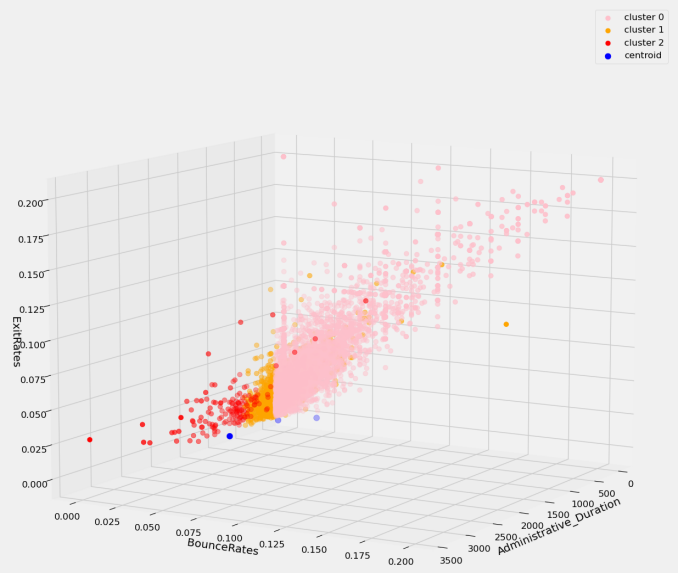
- Precision would tell us cases where the customer did not purchase on the website, but we predicted it as purchased.
- Recall would tell us cases where the customer purchased on the website, but we predicted it as not purchased.

In our case, the recall metric will be more important than precision. Therefore, choosing recall and f1-score, the harmonic mean of both precision and recall, will be more reasonable for evaluation metrics, particularly for class 1.

Furthermore, the AUC-ROC curve is a benchmarking tool for classification problems at various threshold levels. AUC represents the degree or measure of separability, while ROC is a probability curve. It indicates how well the model can distinguish between classes. The greater the AUC, the more accurately the model predicts 0s as 0s and 1s as 1s. By analogy, the higher the AUC, the better the model distinguishes between people who made a purchase and people who did not purchase on the website (Narkhede, 2021).

B. Clustering:

1. K-Means:

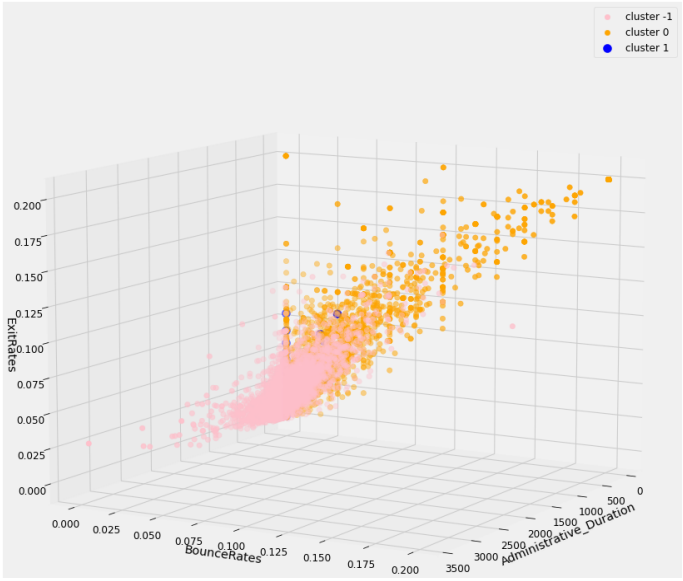


From the responsive plot base on the 3D scatterplot, which defines the algorithm centroid for calculating customer outcome prediction based on the customers' behavior by three set variables, we can see that:

- Cluster-ID 0: customers with the least duration in administrative pages, bounce and exit rates higher.
- Cluster-ID 1: customers that spent medium duration and bounce rate and exit rate medium.
- Cluster-ID 2: customers that spent more time on administrative pages had less chance of exiting the website.

From the cluster behavior, we see that the more time customers spend on administrative-type pages, the less likely they change websites or exit from the website.

2. DBSCAN:



From the 3D scatterplot, group "-1" represents outliers in our model. In this model, we want to analyze three columns "Adminstrative_Duration", "ExitRates" and "BounceRates". In column "Administrative_Duration", most of customers spend less time on this page, which are the outliers, leading to high bounce/exit rates.

V. Conclusion:

A. Classification:

Based on our evaluation metric, the scores of the models we tried are below:

Models	Recall Score for Class 1 (%)	f1-score for Class 1 (%)	ROC AUC (%)	Accuracy (%)
1. Logistic Regression	34	47	50	88.1
2. Support Vector Machine	1	2	50	84.6

Logistic Regression gives better measures against others.

B. Clustering:

There are 12,330 sessions in this dataset. This can be considered a small dataset. Besides, there are outliers, so that we choose **DBSCAN** rather than **K-Means** approach to model. **DBSCAN** groups the customers into small parts based on the outliers, so that we can focus on the target customers better than **K-Means**.

C. Recommendation:

- Improve services and interactions between customers and websites to keep the users stay longer on the website, which increases the chance of purchasing products. For example: mini games to give discount codes, livestream.
- Increase related products so that customers can have more choices.

Reference

Sakar, C.O. *et al.* (2018) *Real-time prediction of online shoppers' purchasing intention using Multilayer Perceptron and LSTM recurrent neural networks - neural computing and applications*, *SpringerLink*. Springer London. Available at: https://link.springer.com/article/10.1007/s00521-018-3523-0 (Accessed: January 19, 2023).

Narkhede, S. (2021) *Understanding AUC - roc curve*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Accessed: January 19, 2023).