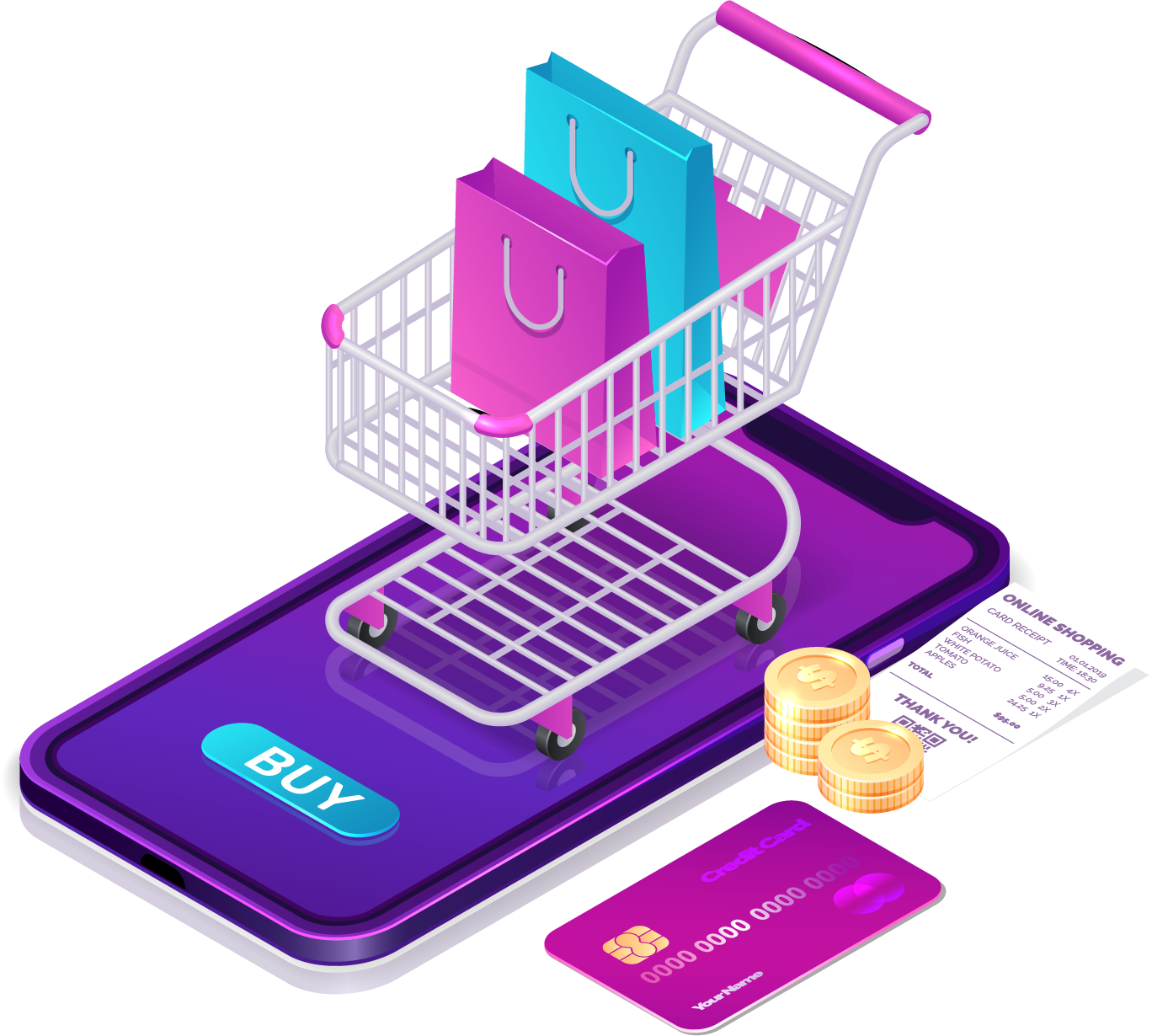
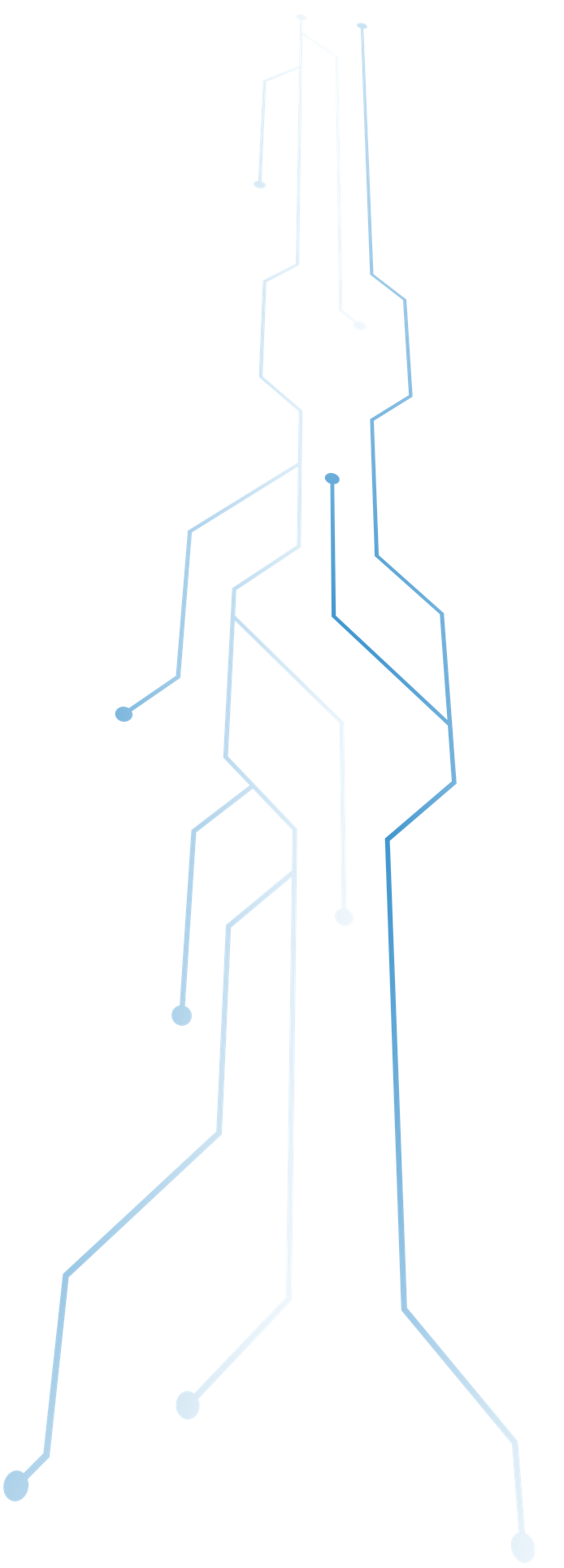
|  |  |
| --- | --- |
| **Submitted By:** | **Mentored By:** |
| Surbhi Bhatia | Mr. Romil Gupta |
| Siddharth Shankar |  |
| Rahul Das |  |
| Vaibhav Bhatnagar |  |
| Jalaj Tripathi |  |

**PGPDSE-FT.G.JUL19 Batch**



Capstone Interim-Project Report

Online Shoppers Purchasing Intention

|  |
| --- |
| **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: 30-10-2019

Place: Gurugram

|  |
| --- |
| **CERTIFICATE OF COMPLETION** |

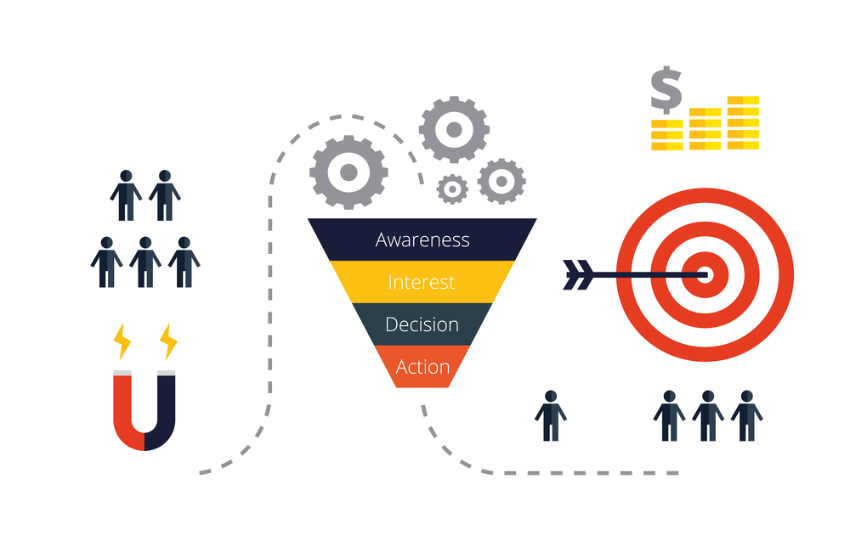
This is to certify that the project titled **“CAPSTONE INTERIM PROJECT REPORT – ONLINE SHOPPERS PURCHASING INTENTION”** was undertaken and completed under the supervision of Mr. Romil Gupta for the Post Graduate Program in Data Science and Engineering (PGPDSE).

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]](#footnote-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](https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset). It is to be noted that the available dataset is a subset of the original data as found in the research paper.[[2]](#footnote-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 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](http://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]](#footnote-3). In Turkey, the Gözalan Group became the distributer of Columbia Sportswear in 2004[[4]](#footnote-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.

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.

*Fig: Current Conversion Rate*

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 |
| 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 and Box-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 and Weekend were One-Hot Encoded. | Month and Revenue features were Label Encoded. |
| Standardization | We apply zero-mean, unit variance normalization. | |
| 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 our next step.

**Visitor Types vs Revenue**

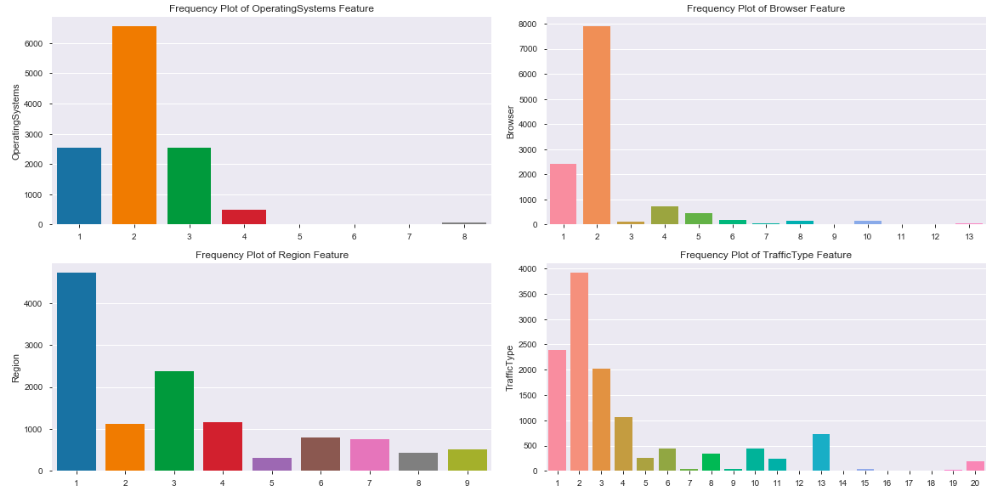
There are 3 different types of visitors: Returning Visitors, New Visitors and Others.

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

|  |  |  |
| --- | --- | --- |
| Visitor Type | Revenue Generated | Overall Percentage |
| New Visitor | False | 75.07% |
| True | **24.92%** |
| Other | False | 80.24% |
| True | 19.75% |
| Returning Visitor | False | 85.90% |
| True | **14.09%** |

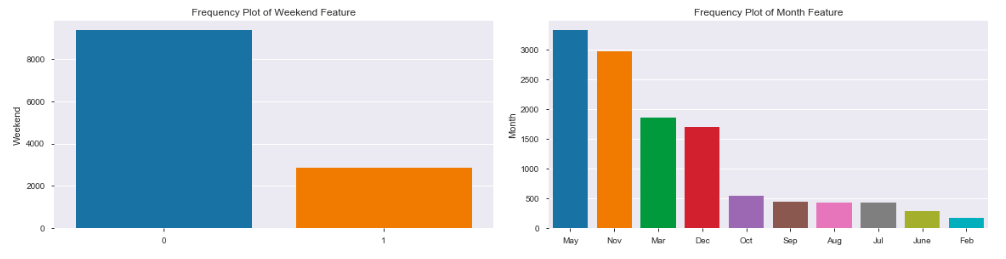
The conversion rate of new visitors is (24.92 – 14.09) ~ 10% greater than returning visitors.

**Below is the bar plot of the following categorical features: 'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'Weekend', 'Month'.**



**Observations:**

* Mostly used OSes are: 2, 3 & 1.
* Mostly used browsers are: 2 & 1.
* Majority of customers are from region 1 & 3.
* Traffic type is mostly via: 1, 2, 3, 4& 13.



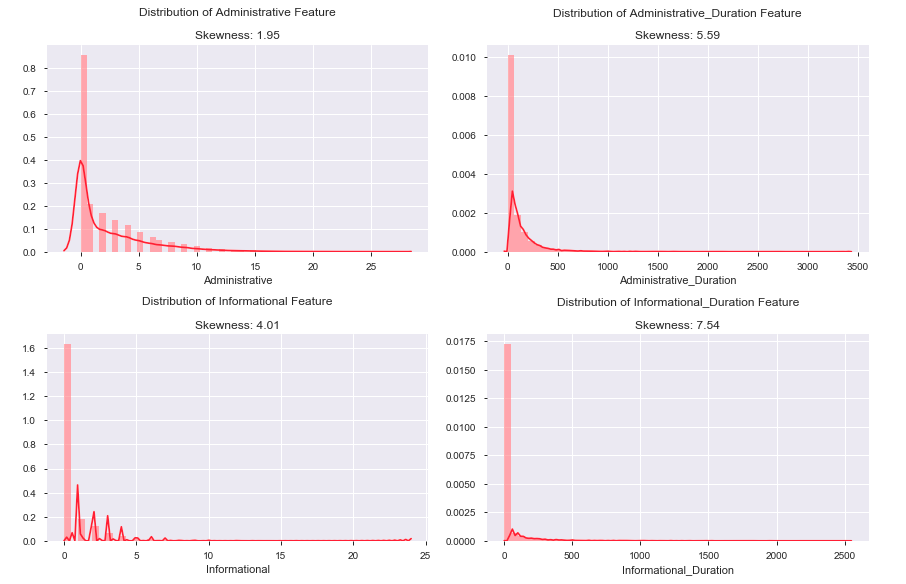
**Observations:**

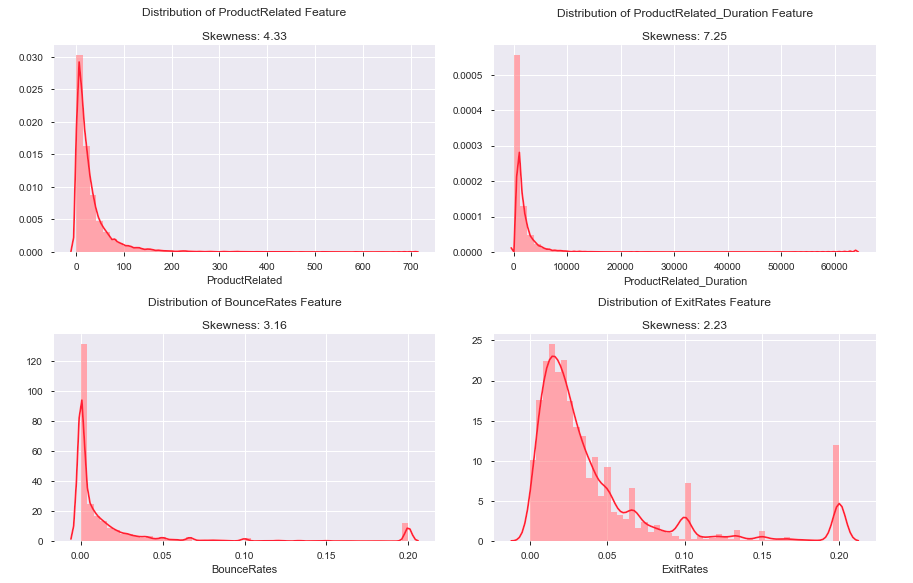
* People mostly visit the site on weekdays.
* Most of the visits has been done in the months of May, November, March and December.

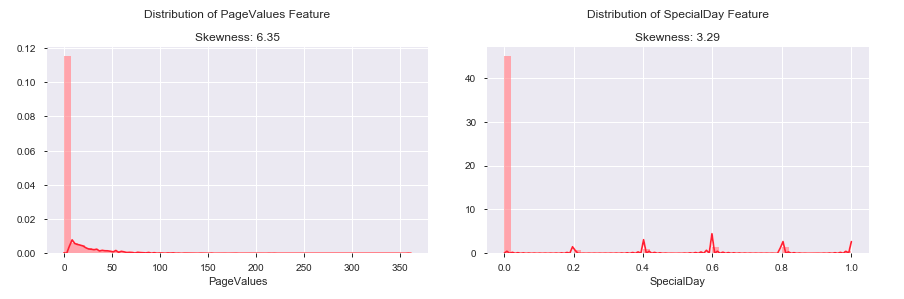
|  |  |
| --- | --- |
| Page Types | Average Time Spent |
| Administrative | 81.65 |
| Informational | 34.83 |
| Product Related | 1206.98 |

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'.**





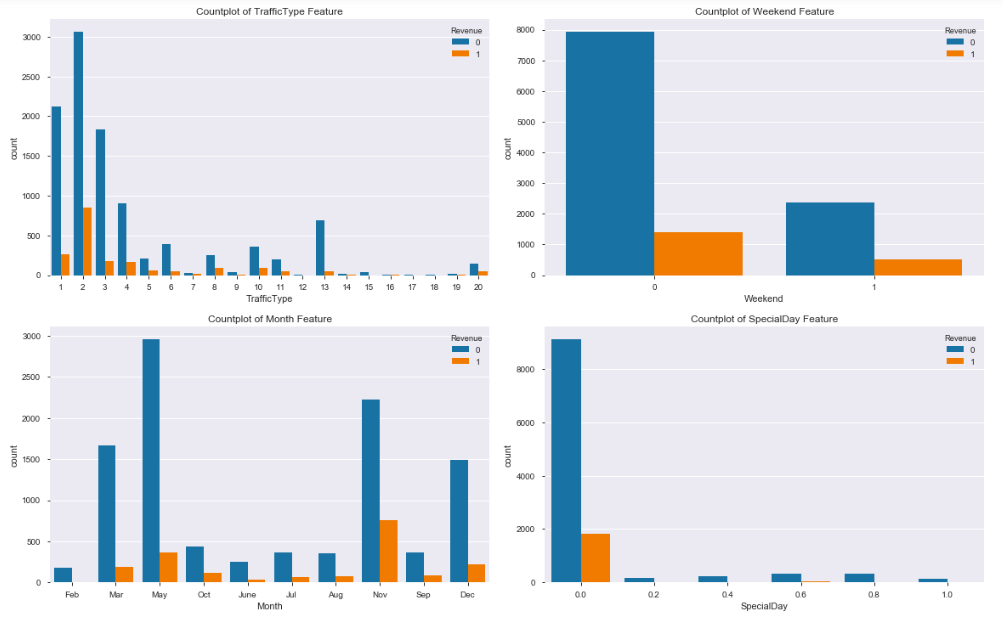
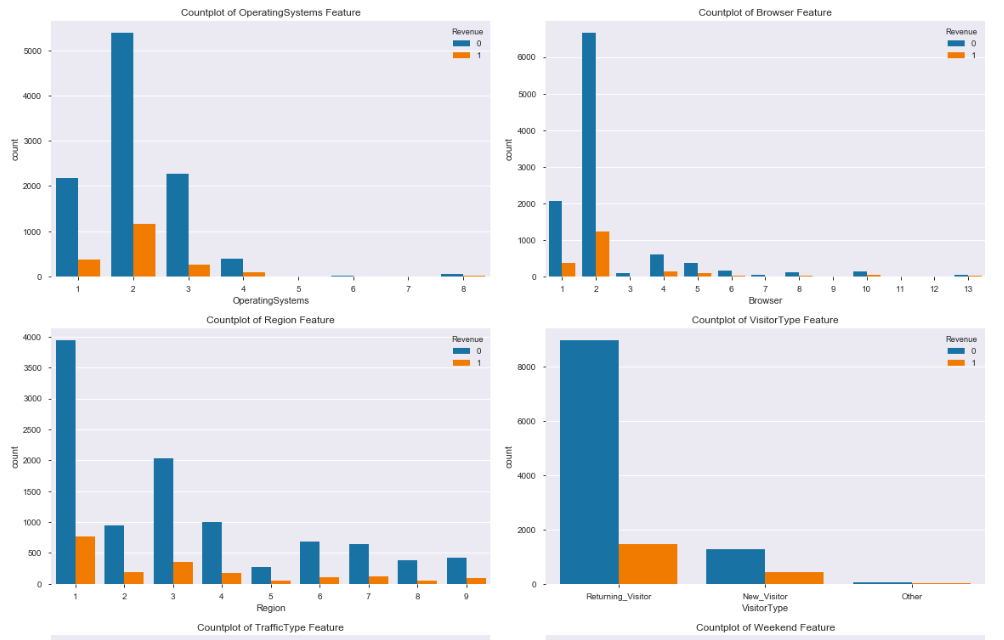


We created a table having the original skewness values of the numerical features and did a comparative analysis by applying log, square root, cube root, 4th power, and 5th power transformations and applied the one having the minimum skewness values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Features | Skewness | Log\_T | Sqrt\_T | Cbrt\_T | 4thPower\_T | 5thPower\_T |
| Administrative | 1.95 | 0.55 | 0.63 | 0.24 | 0.08 | 0 |
| Administrative Duration | 5.59 | 0.24 | 1.53 | 0.69 | 0.37 | 0.21 |
| Informational | 4.01 | 2.01 | 1.93 | 1.62 | 1.51 | 1.46 |
| Informational Duration | 7.54 | 1.9 | 3.42 | 2.43 | 2.05 | 1.87 |
| ProductRelated | 4.33 | -0.04 | 1.5 | 0.81 | 0.45 | 0.19 |
| ProductRelated Duration | 7.25 | -1.43 | 1.41 | 0.42 | -0.21 | -0.72 |
| BounceRates | 3.16 | 3.08 | 1.72 | 0.9 | 0.48 | 0.26 |
| ExitRates | 2.23 | 2.13 | 1.21 | 0.69 | 0.26 | -0.17 |
| PageValues | 6.35 | 1.79 | 2.52 | 1.89 | 1.66 | 1.54 |
| SpecialDay | 3.29 | 3.1 | 2.86 | 2.74 | 2.7 | 2.67 |

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

**Bivariate Analysis:**

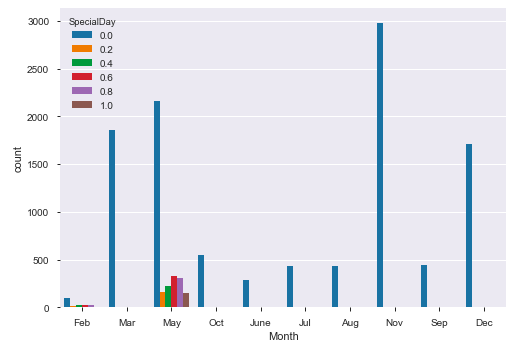
****

**Observations:**

* Revenue is more on Weekdays.
* November has contributed the most in revenue followed by May & December while February followed by June has contributed the least.
* More revenue is generated from the Returning visitors.
* OS Type '2' has generated the most revenue followed by Type '1'.
* Browser Type '2' has generated the most revenue followed by Type '1'.
* Traffic Type '2' has generated the most revenue followed by Type '1' and '3'.
* Majority of revenue are from Region 1, 3 & 4.

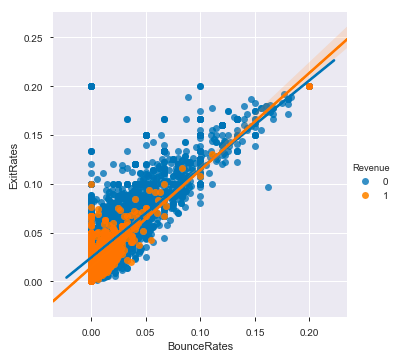
**Special Day vs Month**

We created a count plot of Month, separating data points on Special Day.

****

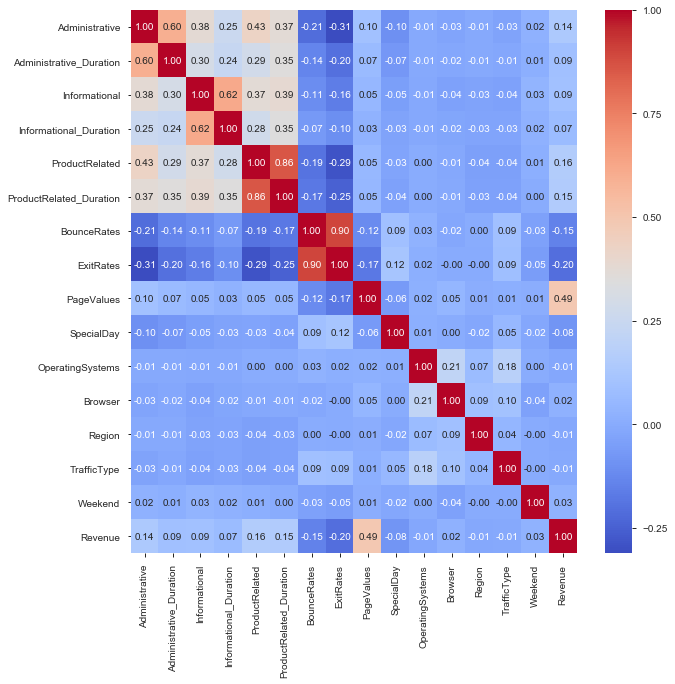
Most of the special days are in the month of May followed by February.

**Exit Rates vs Bounce Rates Correlation:**



Exit Rates and Bounce Rates are highly positively correlated with the correlation coefficient of 0.90.

**Correlation Matrix Using Heatmap:**



**Observations:**

* BounceRates and ExitRates are highly correlated
* PageValues has the highest correlation with the target column (Revenue)
* ProductRelated and ProductRelated\_Duration are highly correlated
* Information and Informational\_Duration are highly correlated
* Administrative and Administrative\_Duration are highly correlated

|  |
| --- |
| **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 :**

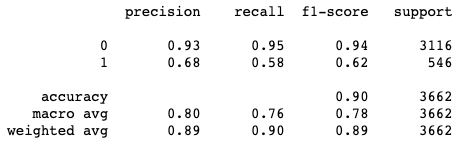
There are a number of metrics used in machine learning tasks and it can be difficult to know which one to choose. Most of the time it will depend on the particular problem and if you have a specific goal to optimize for. Rather than calculating multiple metrics and trying to determine how important each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

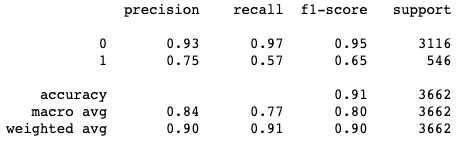
**Base Model:**

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:



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:



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!

Further we will be implementing several machine learning methods, selecting the best model, and optimizing it for our problem using hyperparameter tuning with cross validation.

1. <http://forbes.com/sites/michellegrant/2018/08/14/e-commerce-set-for-global-domination/> [↑](#footnote-ref-1)
2. <https://doi.org/10.1007/s00521-018-3523-0> [↑](#footnote-ref-2)
3. <https://en.wikipedia.org/wiki/Columbia_Sportswear> [↑](#footnote-ref-3)
4. <https://www.nebim.com.tr/en/columbia> [↑](#footnote-ref-4)