**A RESEARCH SYNOPSIS**

**ON**

**ONLINE SHOPPER PURCHASING INTENTION ADOPTED BY**

**GREAT LEARNING**

SUBMITTED FOR THE PARTIAL FULFILLMENT TOWARDS THE AWARD OF POST-GRADUATION IN DATASCIENCE ENGINEERING, POWAI- MUMBAI

**SUBMITTED BY** PRIYA CHANDAK TANU KHATIYAN SIDDHI MAURYA

SHASHANK SURESHKUMAR AKSHAY PATIL

**UNDER THE GUIDENCE OF**

MR.DIPANJAN GOSWAMI

1. Introduction……………………………………………………………………………………………………………………………………………………………….3

[1.1 ONLINE SHOPPER PURCHASING INTENSION 3](#_Toc28881362)

[1.2 AIM OF THE PROJECT 3](#_Toc28881363)

[1.3 PROBLEM STATEMENT: 3](#_Toc28881364)

[2. LITERATURE SUMMARY 4](#_Toc28881365)

[2.1. HANDLE IMBALACED CLASSIFICATION PROBLEMS IN MACHINE LEARNING 4](#_Toc28881366)

[2.2. ALGORITHMS FOR SCREENING OF ONLINE SHOPPER PURCHASING INTENTION 4](#_Toc28881367)

[3. DATA DESCRIPTION 5](#_Toc28881368)

[3.1 DATA SET: 5](#_Toc28881369)

[3.2 VARIABLES CONSIDERED FOR ANALYSIS 7](#_Toc28881370)

[3.3 TARGET VARIABLES: 8](#_Toc28881371)

[4.1 INTRODUCTION: 9](#_Toc28881372)

[5A. DATA CLEANING 17](#_Toc28881374)

[5B. Clustering Analysis 20](#_Toc28881375)

[6. ARCHITECTURE 22](#_Toc28881376)

[6.1 HANDLING IMBALANCED DATA 23](#_Toc28881377)

[7. TENTATIVE LIST OF ALGORITHMS & INITIAL APPROACH 25](#_Toc28881378)

[7.1 LOGISTIC REGRESSION 25](#_Toc28881379)

[7.2 DECISION TREE (CART) 26](#_Toc28881380)

[Layout / flow of Decision Tree 27](#_Toc28881381)

[Disadvantages of CART 27](#_Toc28881382)

[7.3 RANDOM FOREST 28](#_Toc28881383)

[Advantages of Random Forest 28](#_Toc28881384)

[Disadvantages of Random Forest 28](#_Toc28881385)

[8. INITIAL APPROACH 29](#_Toc28881386)

9. Implementation………………………………………………………………………………………………………………………..………………31

[10. RESULTS AND COMPARISON STUDY 35](#_Toc28881387)

**1. INTRODUCTION**

## 1.1 ONLINE SHOPPER PURCHASING INTENSION

Online shopping sites are fast replacing traditional or physical shops. Over the years, the trust of the customers for online shopping sites has increased considerably. The increase in the number of these sites, on one hand, has led to a fierce competition, which means better and cheaper products for customers. When consumers have intention to purchase then they gather information, make comparison, evaluation and take decision. Consumer purchase intention is an important predictor for online shopping if consumer have intention then they can behave. The purpose of this study is to determine the consumer purchase intention influence on online shopping behavior with the moderating role of attitude. Global trends show that people moving towards online shopping rapidly it is a third most popular activity in whole world. The online shopper purchasing intension analysis system consisting of two modules which simultaneously predicts the visitor’s shopping intention and Web site abandonment. In the first module, we predict the purchasing intention of the visitor using dataset kept track during the visit along with some session and user information.

## 1.2 AIM OF THE PROJECT

The purpose of current research is to examine the influence of consumer purchase intention on online shopping behavior and trying to make stronger this relationship so that people can actual behavior. We aim to predict these online shoppers purchasing intension with a reasonable level of accuracy using collected shopping data and aid in the early detection and subsequent treatment.

## 1.3 PROBLEM STATEMENT:

* Here we will predict whether user will complete the transaction or not based on other few dependent variables.
* In case of online shopping it is important that the sale of product needs to be forecasted accurately.
* If we forecast wrong prediction of user type (new user, existing user or other user) purchase and website page experience than our online sale website may get less rating and seek customer to other competitor website. Our data set is related to page wise session and user type, browser type traffic coming on our website.
* So, it’s necessary to forecast the correct user experience on web pages and product purchase ratio.

The task is to predict the user experience on website page and predict the Revenue and overall website user experience.

# 2. LITERATURE SUMMARY

## HANDLE IMBALACED CLASSIFICATION PROBLEMS IN MACHINE LEARNING

While performing the machine learning or any operation on the data that is imbalanced the model will be inaccurate and biased. This problem is predominant in scenarios where anomaly detection is crucial for data. Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. This happens because Machine Learning Algorithms

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>

## ALGORITHMS FOR SCREENING OF ONLINE SHOPPER PURCHASING INTENTION

There are various algorithms and methodologies used for online shopper purchase intention by segmenting and classifying Revenue into different categories. An attempt has been made to furnish the reader with an insight of Machine Learning algorithms like k- NN (k-Nearest Neighbors), Logistic Regression, RFT (Random Forest Trees), CART (Classification and Regression Trees) and Hierarchical clustering algorithm for feature extraction, Revenue segmentation and classification.

# 3. DATA DESCRIPTION

## DATA SET:

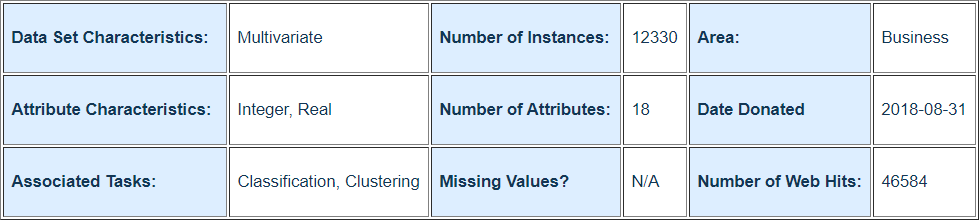
This demo project is based on an Online Shoppers Purchasing Intention dataset on UCI Machine Learning Repository, which can be used to predict whether users will complete their transactions, a binary classification task. This dataset was first published in an article from 2018. In the article, the authors compared random forest, support vector machines, and multi-layer perceptron’s on this data, and developed a prediction model using multi-layer perceptron with selected variables for predicting the purchasing intention of users. The model is able to help e-commerce businesses identify customers who are more likely to complete transactions, and adjust marketing strategies accordingly.

The dataset consists of 12330 records, each containing metrics of web visits of a user within a one-year timeframe. 85.4% (10422) of the customers did not complete the transaction. Positive examples, i.e. those who completed transactions, only take up 15.5% (1908) of the dataset. There are both continuous and categorical features in the data. Continuous variables include the total number of visits and time spent in three types of websites. There is also the average Google Analytics web metrics for all the websites that the users visit. Categorical variables record the user profiles and session information, including the type of operation systems and browsers, time and locations of the visits, etc.

The dataset, “Online shopper purchasing intention” was obtained from the UCI Repository. The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different userina1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

The data was collected by C.Okan Sarkar Department of Computer Engineering, Faculty of Engineering and Natural Science, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey. Yomi Kastro Inveon Information Technologies Consultancy and Trade 34335 Istanbul, Turkey. Dataset contains "Administrative", "Administrative Duration", "Informational", "Informational Duration", and “Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories information, and shopper purchasing intention history of 12330 records from Turkey. There are no missing values in this dataset; the dataset consists of 10 numeric, with 8 categorical attributes.

([https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset](https://archive.ics.uci.edu/ml/datasets/Online%2BShoppers%2BPurchasing%2BIntention%2BDataset))



## VARIABLES CONSIDERED FOR ANALYSIS

* + - The dataset consists of 10 numerical and 8 categorical attributes.
    - The 'Revenue' attribute can be used as the class label

*Below is the detailed description of each of the variables****.***

|  |  |  |
| --- | --- | --- |
| **variable** | **Data Type** | **Description** |
| Administrative | Int64 | Administrative represent the number of different types of  pages visited by the visitor in that session and total time  spent in each of these page categories administrative  Value. |
| Administrative\_Duration | float64 | Administrative\_Duration represents the duration taken by  User to takes an action, e.g. moving from one page to another administrative Page. |
| Informational | Int64 | Informational represents the Informational Value |
| Informational\_Duration | float64 | Duration in Informational Page |
| ProductRelated | Int64 | Product Related Value |
| ProductRelated\_Duration | float64 | Duration in Product Related Page |
| BounceRates | float64 | Feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. |
| ExitRates | float64 | Feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. |
| PageValues | float64 | Feature represents the average value for a web page that  A user visited before completing an e-commerce transaction. |
| SpecialDay | float64 | Feature represents the average value for a web page that a user visited before completing an e-commerce transaction. |
| Month | object | Feature indicates the closeness of the site visiting time to  a specific special day (e.g. Mother’s Day, Valentine's Day)  in which the sessions are more likely to be finalized with  Transaction. The value of this attribute is determined by  Considering the dynamics of e-commerce such as the duration between the order date and delivery date. |

|  |  |  |
| --- | --- | --- |
| OperatingSystems | Int64 | Operating system used |
| Browser | Int64 | Browser used |
| Region | Int64 | Region of the user |
| TrafficType | Int64 | Traffic Type |
| VisitorType | object | Types of Visitor |
| Weekend | bool | Weekend or not |
| Revenue | bool | Revenue will be generated or not |

## TARGET VARIABLES:

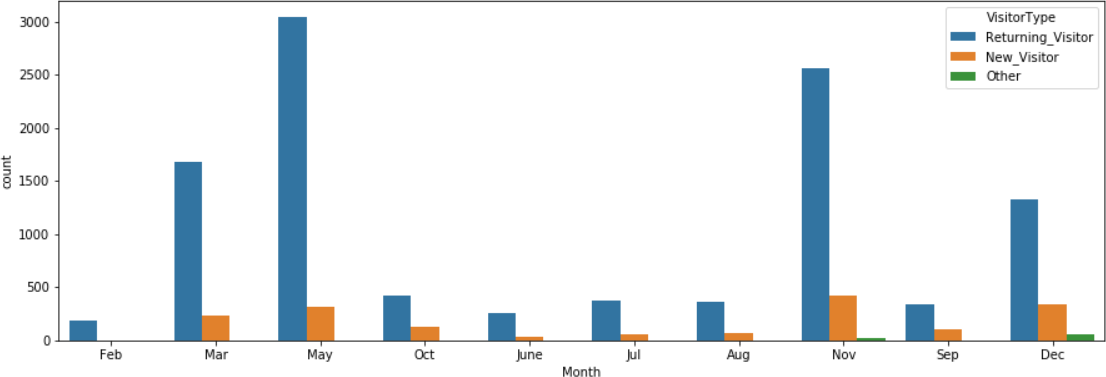
The target variable is “Revenue” and that is characterized by ‘TRUE’ or ‘FALSE’ in our data set where ‘TRUE’ represents that customer will complete the transaction of the product and ‘FALSE’ indicates that customer will not complete the transaction of that product.

**4. EXPLORATORY DATA ANALYTICS**

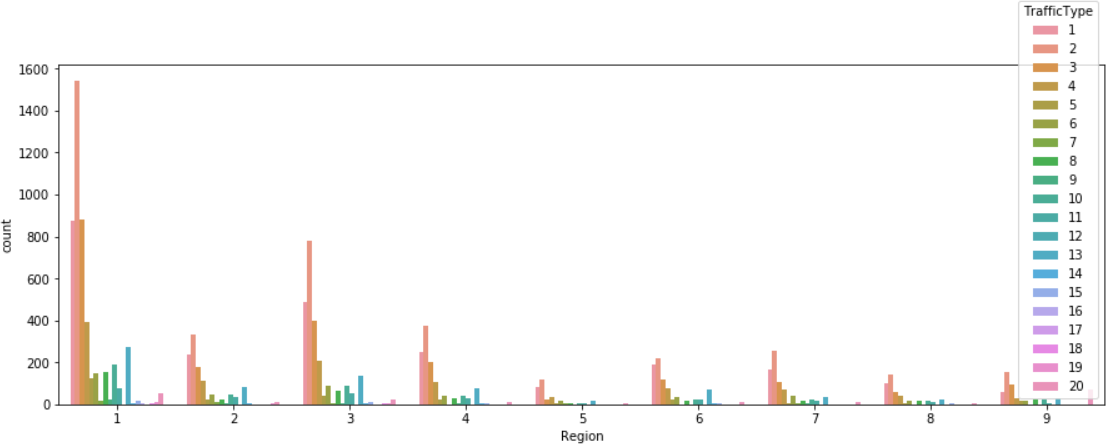
## INTRODUCTION:

EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data. The target variable for our study is ‘Revenue’ and as it is a categorical variable, we use Regression Analysis.

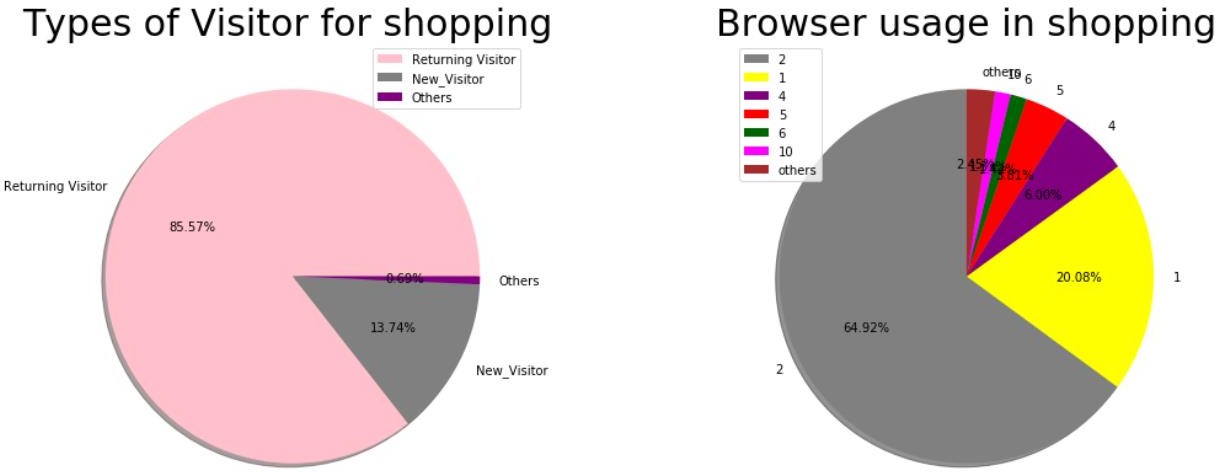
#### BIVARIATE ANALYSIS



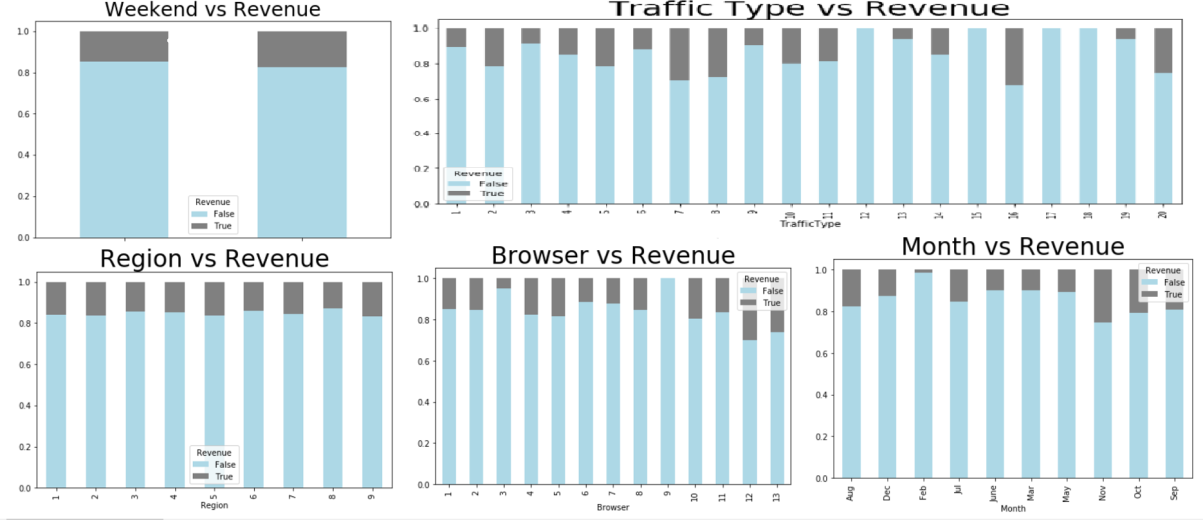
Here we have May have high returning\_visitor and in Nov most of new\_visitor is arriving others in Dec.

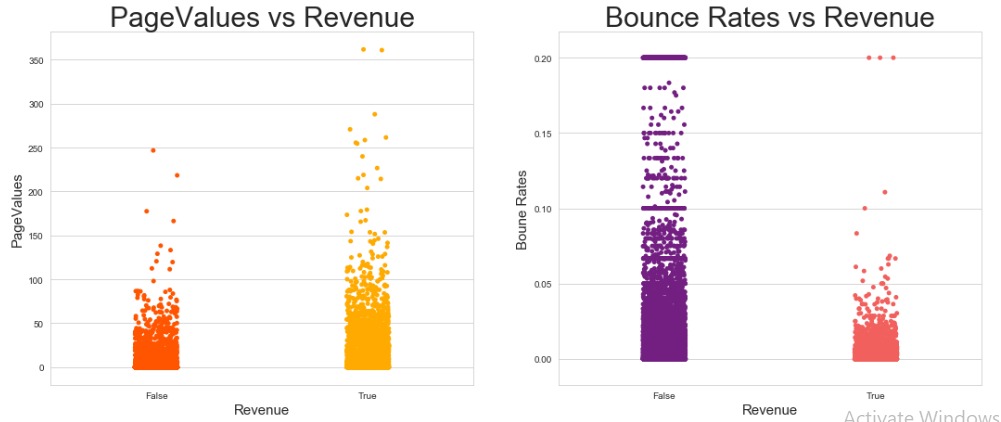


### In region1 Traffic Type is high, which means that region1 has more chance to get revenue.

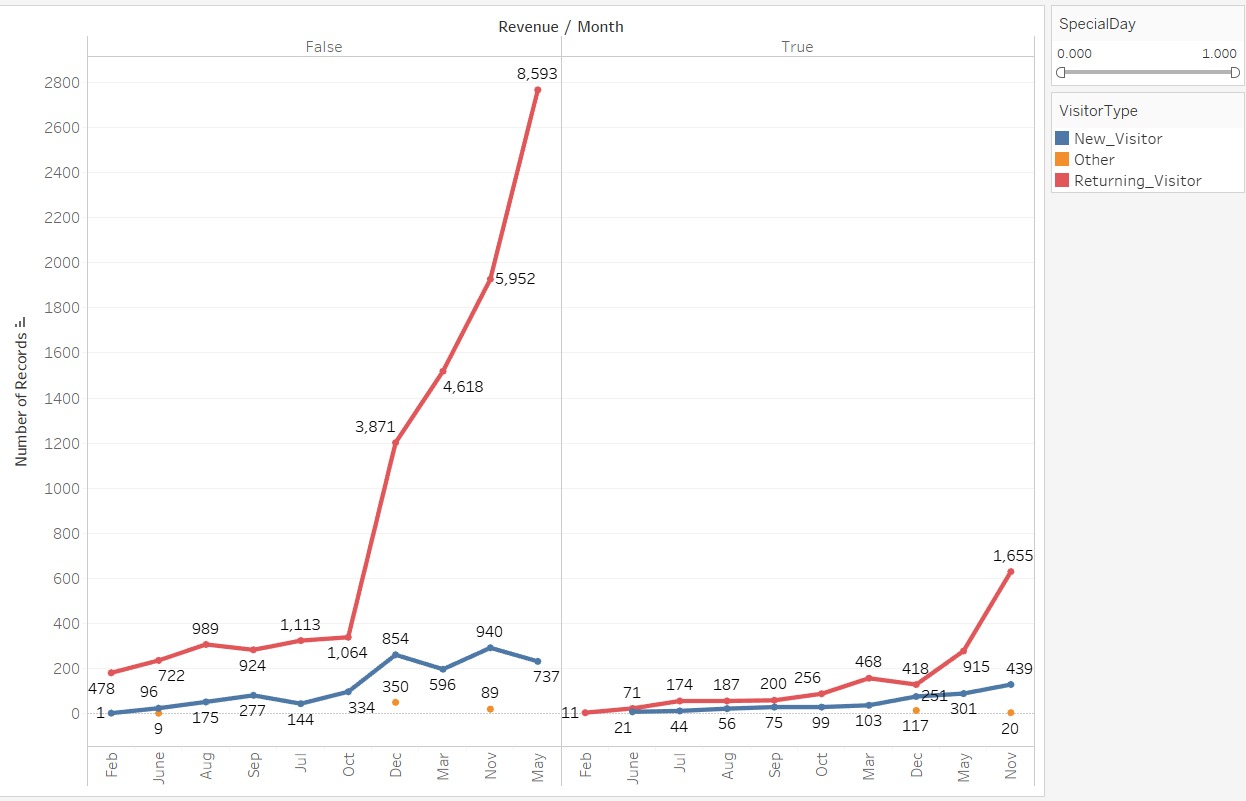


Here in VisitorType Returning visitor is more than any other visitor and in browser usage is high in browser2 which indicates that here we have chance to get revenue.

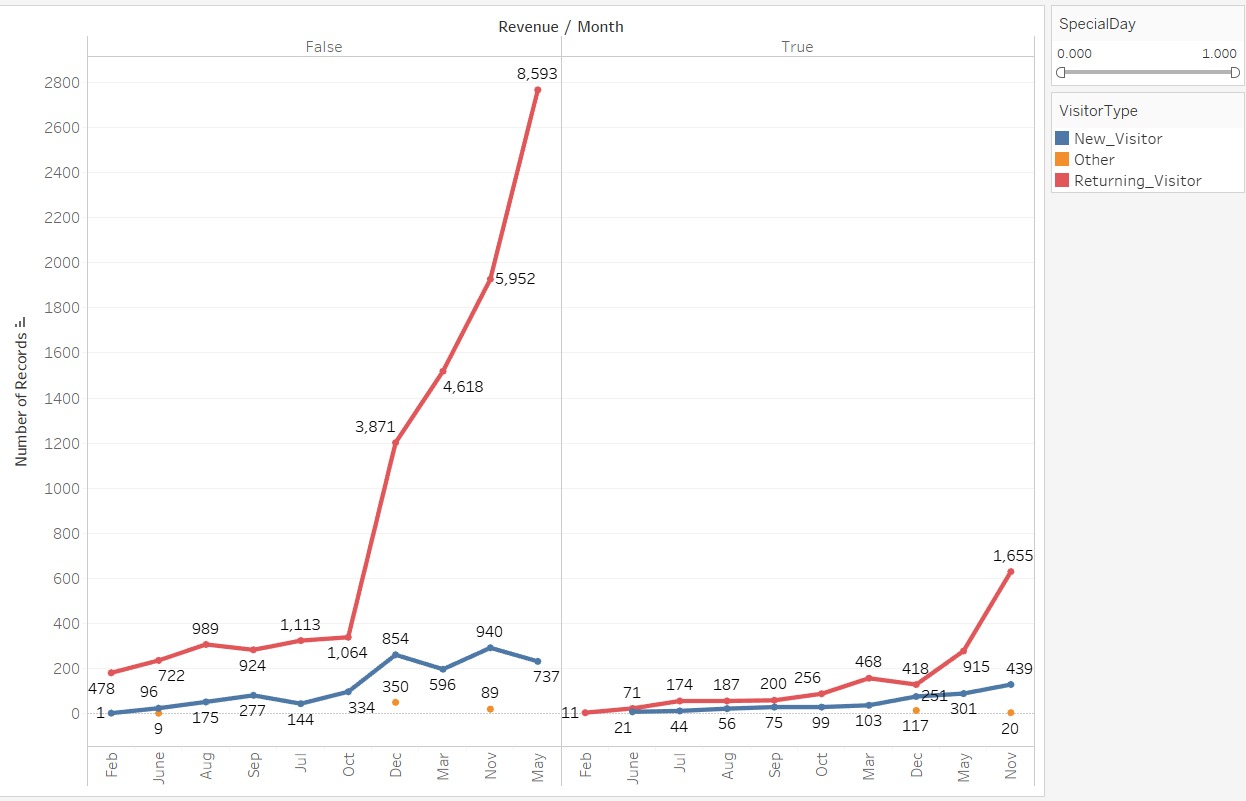




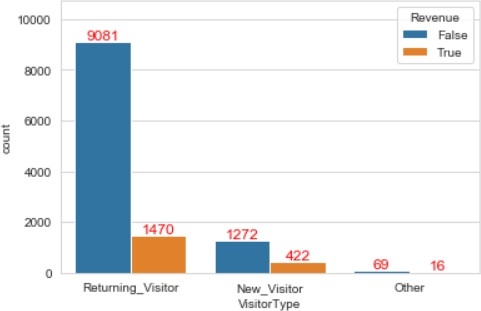
* If the page value is more then there is more chance that customer will contribute in revenue.
* Where the bounce rate is more there is less chance of revenue



Here the graph is showing that browser, region and traffic type has high frequency in returning visitor.

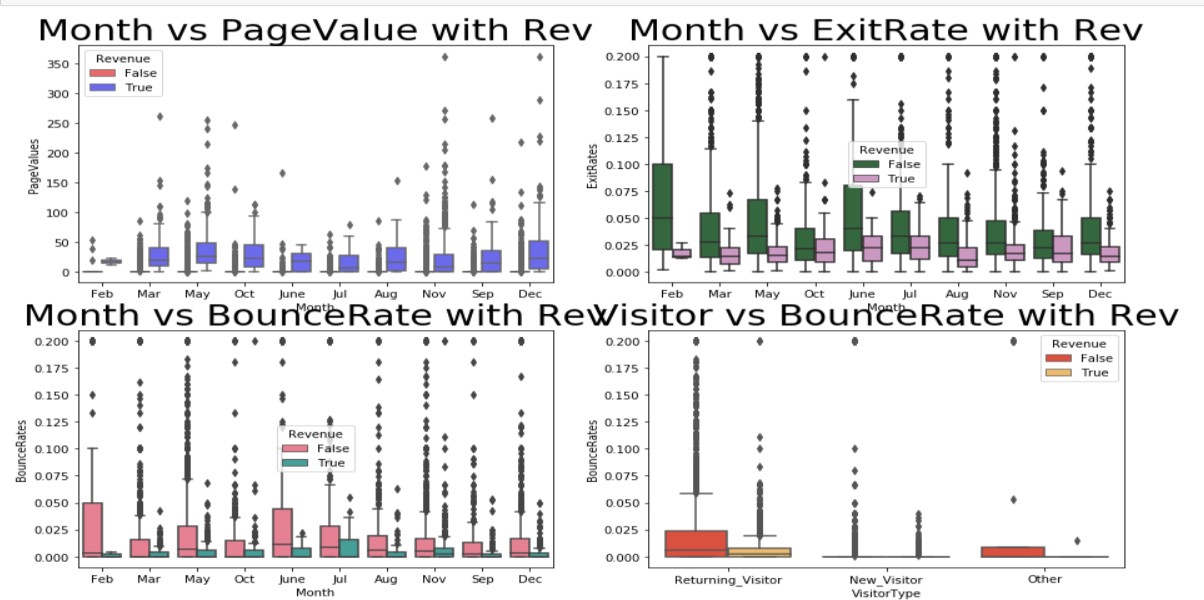


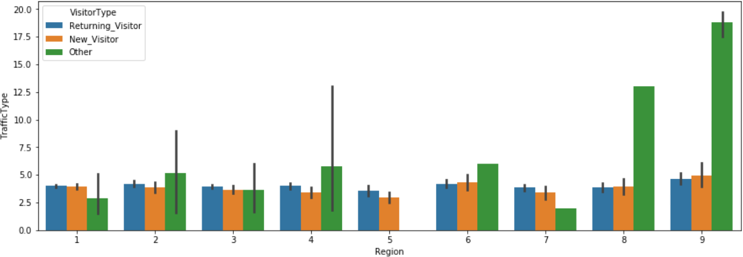
Month vs revenue w.r.t visitor type and special day



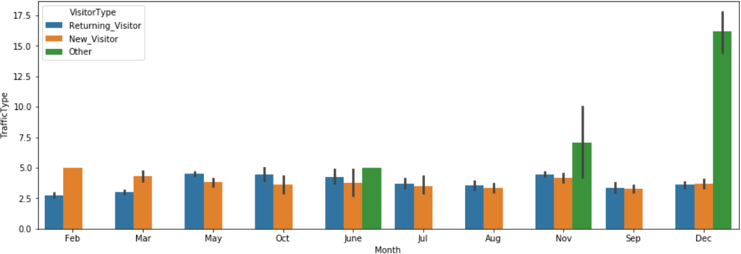
In this plot returning visitor is has high frequency and it has revenue true also has high frequency which means that visitor will spend more time on returning visitor.

#### 4.3 MULTIVARIATE ANALYSIS

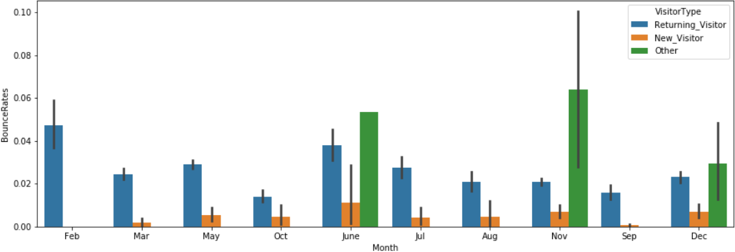




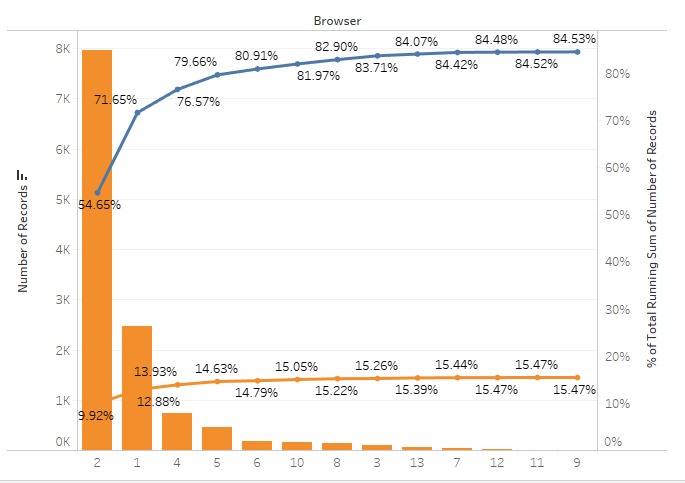
We observe that there is large number of visitors is involved in traffic based on different regions.



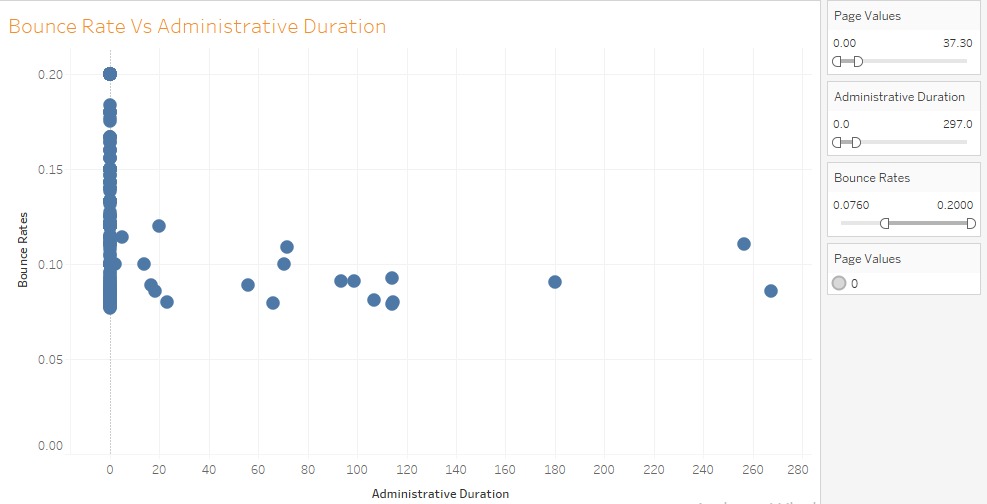
On the monthly basis, returning visitor and new visitor are often seeing in a year. But at end of the year i.e. in the month of December, we observe there is highest number of other visitors.



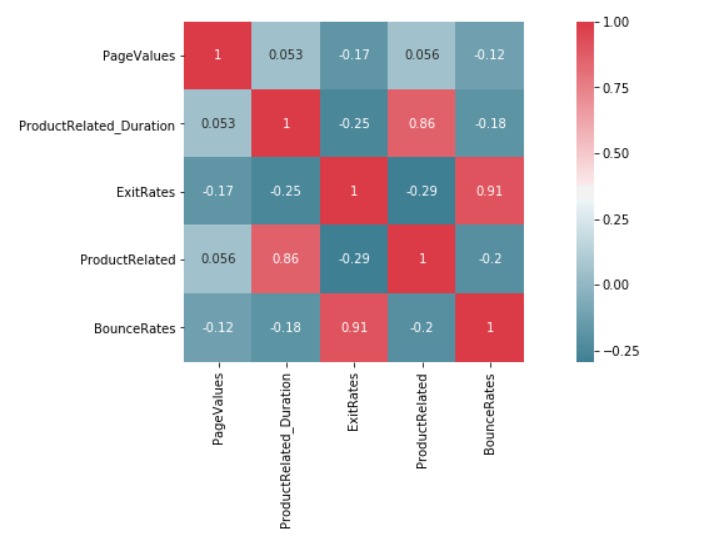
In the each month we can observe there is returning visitor, while new visitor has bounce rates constant. For the other visitor type bounce rate is highest in the month of November.



90% of users are coming through 2 and 1 browser. Out of 15% of converted users 13% we are getting from 1 and 2 browsers.



Where Page value is fewer visitors are not spending time there also bounce rate is also more there.



* Correlation can vary from -1 to +1.
* If correlation is close to +1 indicate that this feature is positively correlated with target variable.
* If correlation is close to -1 indicate that this feature is negatively correlated with target variable.
* Significant columns: Which are highly correlated with the Revenue(target) ['PageValue']
* Redundant columns: Which are highly correlated with the Independent variable:

'ProductRelated' and 'ProductRelated\_Duration'

'BounceRates' and 'ExitRates'

'Administrative' and 'Administrative\_Duration',

'Informational' and 'Informational\_Duration'

# 5A. DATA CLEANING

#### MISSING VALUE TREATMENT

For the categorical column like Revenue, Weekend, Month, Informational etc., the null values were replaced by the **mode**. And null values in the numerical column like Operating system, Browser, region etc. present were replaced by their **median** values.

#### OUTLIER TREATMENTS

Since the data is shopper purchasing intention in nature, outlier treatment is not advisable as each data point is significant and we need to include them as they are in our analysis.

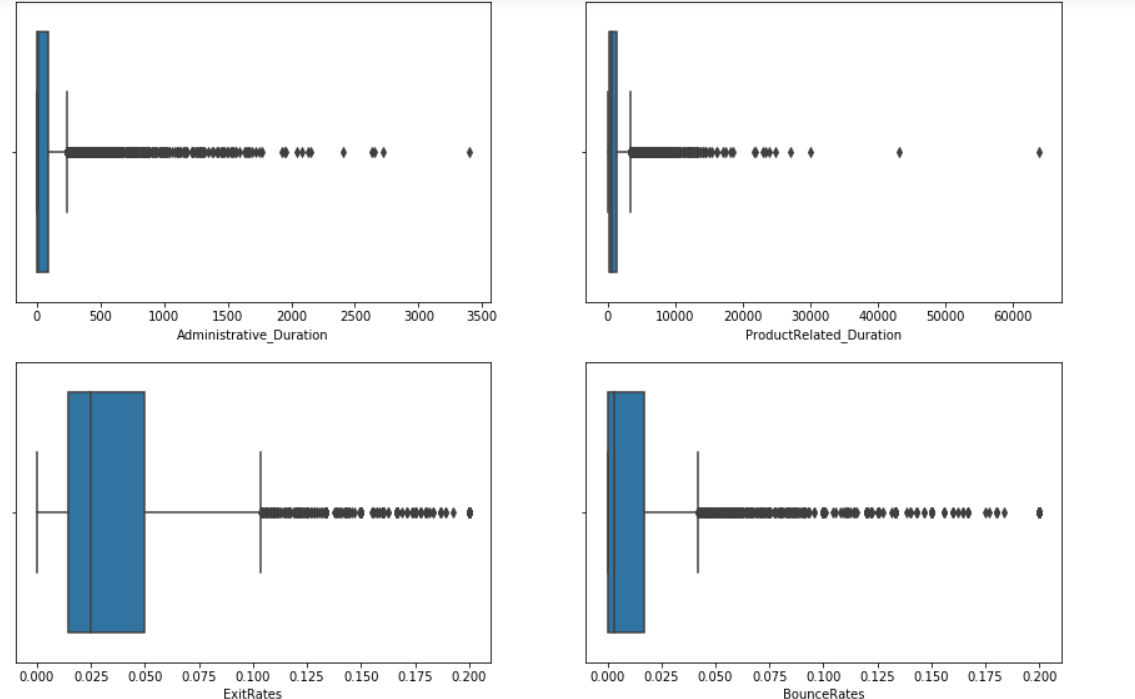
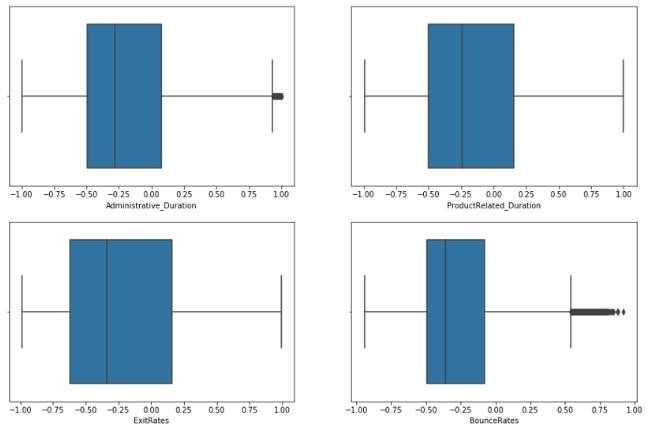


Figure: Dealing with outlier

* + - There are many outliers but we can’t remove the outliers because it may the behavior of the customer, we will lose the info.

#### Observations:

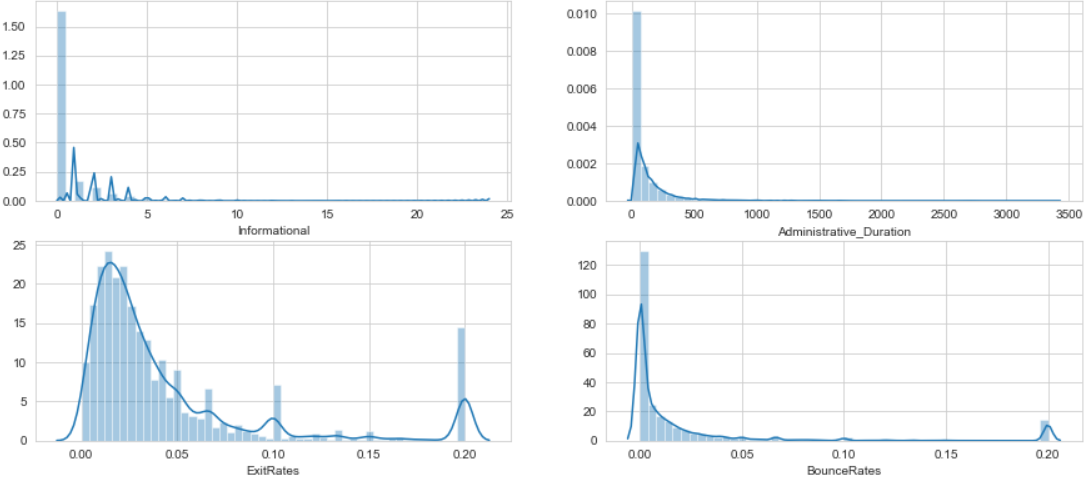
* + - 50% of the users have Administrative\_Duration close to 0.(These are uninterested customers)
    - 25% of the users have the bounce rate 0.(These are interested customers)
    - After normalization most of the outliers has treated.



Compare to previous graph this graph is better it has less outlier value and it removes the all outlier from exit rate and product related duration and others have some outlier.

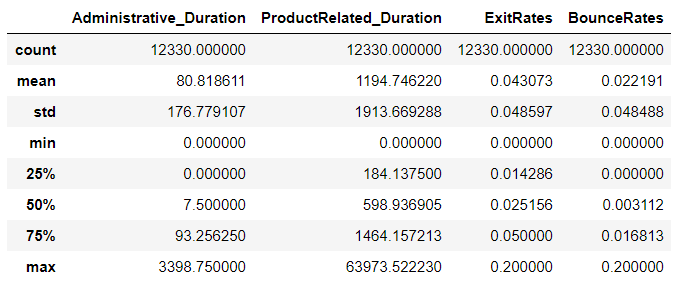
#### 5.3 Normalization

We will check whether our variable is normally distributed or not

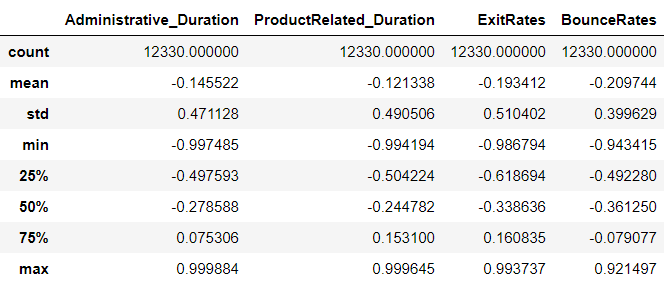


We observe that some of the features are right skewed, we can apply transformation to handle the right skewed data.

Before Normalization-

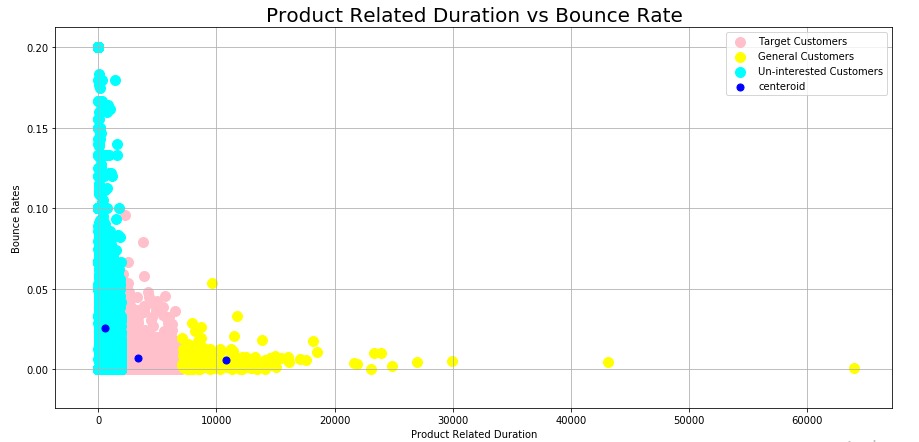


After Normalization-



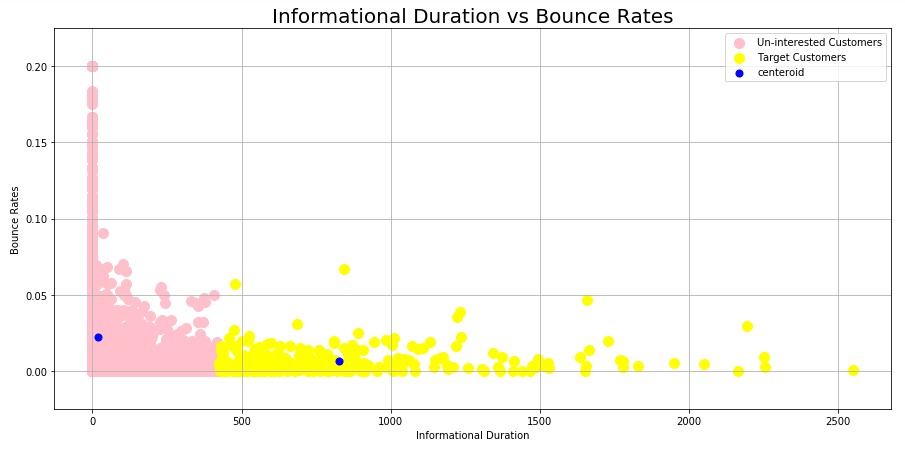
# 5B. Clustering Analysis

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.



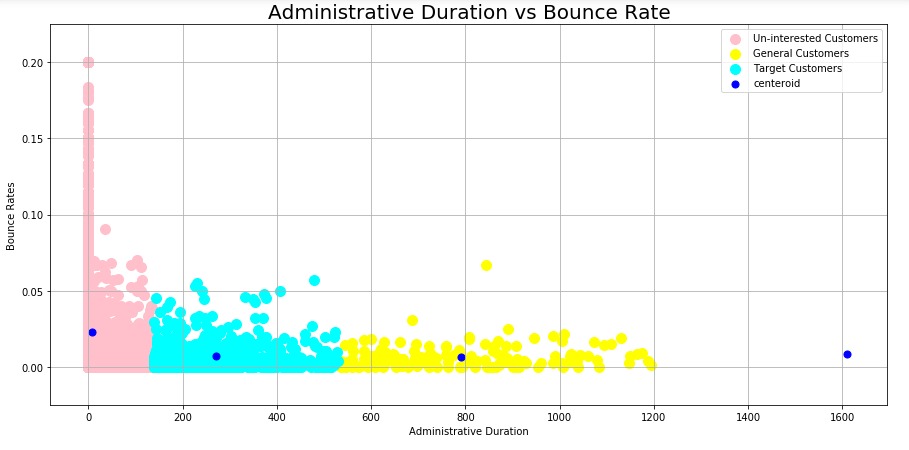
By, looking at this Clustering plot, we can say confidently that the customers who spent a longer Product Related duration in a website are very less likely to bounce from the website that is navigating away from the website just after navigating one page of that website.

There are Three Groups, The cyan Group is a group of customers who stay for shortest product duration and have highest chance for Navigating away from a website



By, looking at this Clustering plot, we can say that the customers who spent a longer Informational duration in a website are very less likely to bounce from the website that is navigating away from the website just after navigating one page of that website.

There are Two Groups, The Pink Group is a group of customers who stay for shortest Informational and have highest chance for Navigating away from a website



By, looking at this Clustering plot, we can say that the customers who spent a longer administrative duration in a website are very less likely to bounce from the website that is navigating away from the website just after navigating one page of that website.

There are Three Groups, The Pink Group is a group of customers who stay for shortest administrative duration and have highest chance for Navigating away from a website

# 6. ARCHITECTURE

We have to take a different approach here from a normal machine learning flow because of the nature of our data. The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets.

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards Revenue which has number of instances. They tend to only predict the majority Revenue data. The features of the minority Revenue are treated as noise and are often ignored. Evaluation of a classification algorithm performance is measured by the Confusion Matrix which contains information about the actual and the predicted Revenue.

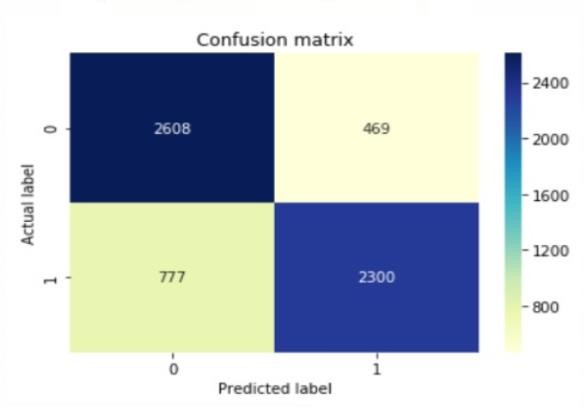
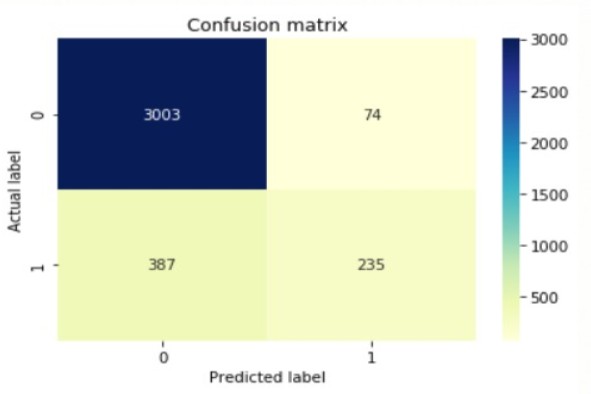


Figure: For base model figure: after feature engineering and smote

|  |  |  |
| --- | --- | --- |
| 0 | TN | FP |
| 1 | FN | TP |
|  | 0 | 1 |

ACTUAL

PREDICTED

However, while working in an imbalanced domain accuracy is not an appropriate measure to evaluate model performance**.** For e.g.: A classifier which achieves an accuracy of 98 % with an event rate of 2

% is not accurate, if it classifies all instances as the majority class. And eliminates the 2 % minority class observations as noise.

To fully evaluate the effectiveness of our model, we must examine **precision** and **recall** as well. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa.

**Precision:** What proportion of positive identifications was actually correct?

**Recall:** What proportion of actual positives was identified correctly?



### HANDLING IMBALANCED DATA

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes.

* 1. RESAMPLING TECHNIQUES
     1. **Random under-sampling**

Random Under sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out.

#### Advantages

* + - * + It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.

#### Disadvantages

* + - * + It can discard potentially useful information which could be important for building rule classifiers.
        + The sample chosen by random under sampling may be a biased sample. And it will not be an accurate representative of the population. Thereby, resulting in inaccurate results with the actual test data set.

#### RANDOM OVER-SAMPLING

Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

#### Advantages

* + - * + Unlike under sampling this method leads to no information loss.
        + Outperforms under sampling

#### Disadvantages

* + - * + It increases the likelihood of overfitting since it replicates the minority class events.

#### SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.

#### Advantages

* + - * + Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances
        + No loss of useful information

#### Disadvantages

* + - * + While generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise
        + SMOTE is not very effective for high dimensional data

# 7. TENTATIVE LIST OF ALGORITHMS & INITIAL APPROACH

Since our problem is a classification problem, we will be using the following algorithms in modelling:

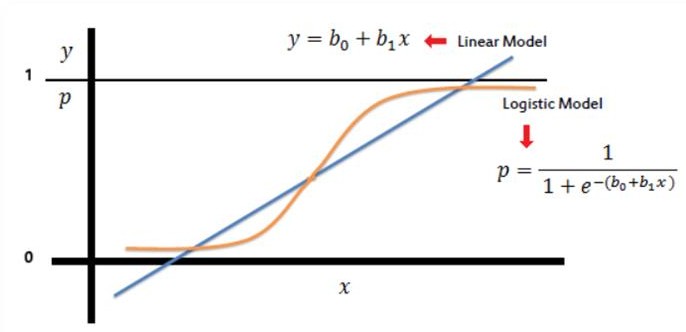
* + - * Logistic Regression
      * Tree Based Classifiers / Regressor
        + Decision Tree
        + Random Forest

## LOGISTIC REGRESSION

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

* + - A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
    - Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



#### Assumptions or Requirements of Logistic Regression:

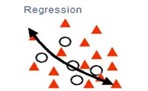
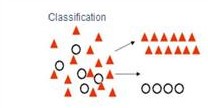
1. First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
2. Second, logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.
3. Third, logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.
4. Fourth, logistic regression assumes linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
5. Finally, logistic regression typically requires a large sample size.

## DECISION TREE (CART)

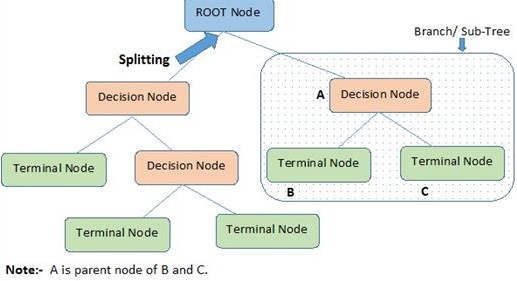
A Decision tree (CART) is a schematic, tree-shaped diagram used to determine a course of action or show a statistical probability. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

#### Types of Decision Tree

* + - **Classification Trees:** where the Dependent variable is categorical and the tree is used to identify the "class" within which a Dependent variable would likely fall into.
    - **Regression Trees:** where the Dependent variable is continuous and tree is used to predict its value. (E.g. the price of a house, or a patient's length of stay in a hospital).



## Layout / flow of Decision Tree



**Advantages of CART**

* Simple to understand, interpret, visualize.
* Decision trees implicitly perform variable screening or feature selection.
* Can handle both numerical and categorical data. Can also handle multi-output problems.
* Decision trees require relatively little effort from users for data preparation.
* Nonlinear relationships between parameters do not affect tree performance.

## Disadvantages of CART

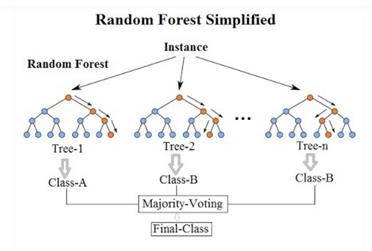
* Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.
* Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting.
* Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.
* Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree.

## RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems.



Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, we don’t have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest repressor.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

## Advantages of Random Forest

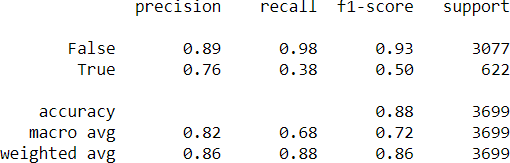
* + - There is no need for feature normalization
    - Individual decision trees can be trained in parallel
    - Reduced overfitting
    - Require almost no input preparation
    - Performs implicit feature selection
    - It’s very quick to train

## Disadvantages of Random Forest

* + - No interpretability

# 8. INITIAL APPROACH

1. Made a Decision Tree model on the entire data to see its overall behavior and check entropy.



While the accuracy of this simple DT was high, it had very low recall and precision as expected because of the class imbalance.

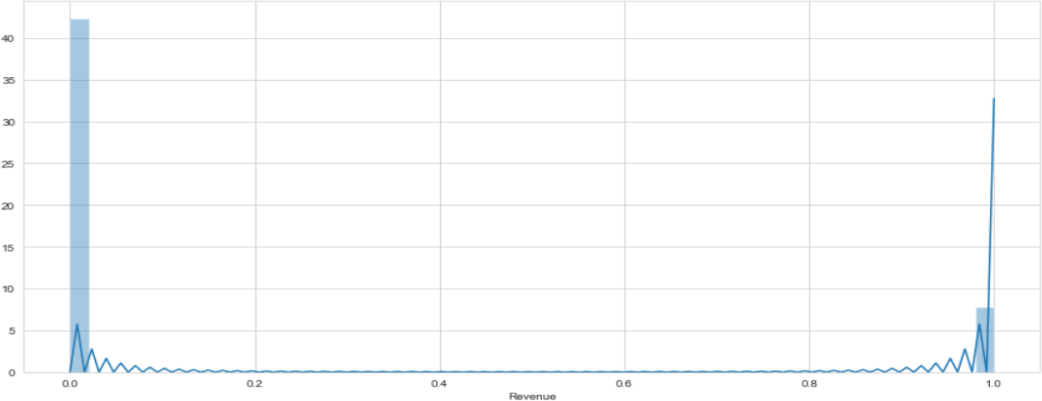
1. Used IBM SPSS® software to get the Co-efficient of variation, coefficient of correlation, skewness and kurtosis of the predictor variables to aid in feature selection on the following guidelines employing the meta features:
   1. Features which have low variance i.e. low coefficient of variation are candidates for elimination.
   2. Features which are relatively unrelated with other features i.e. low average correlation can be eliminated.
   3. Features which have lower entropy i.e. lesser information content can be eliminated.
   4. Features which have highly asymmetric distribution measured by skewness are more suitable to be removed.
   5. Features with exhibit varying peaks measure in terms of kurtosis scan are eliminated.

***This however did not provide us any significant and actionable insights on our dataset.***

1. To find the significant variables we will do statistical tests:

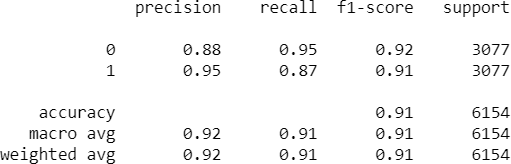
Target variable is Categorical, so we will use two Statistical tests to check the association between Independent Variables and Dependent Variable:

* 1. T-Test
  2. Chi-Square Test



From the above plot we can see that output variable is not normally distributed. We will do statistical test of Shapiro to test normality.

Moreover, the output variable has imbalance data so we need to do SMOTE technique to reduce misbalancing of data.



As we can observe, there was significant improvement in our results.

After the above attempts, the focus was shifted towards finding the most optimal way of feature selection; And for our data set, that was by running a **chi-square test for independence** on each of the individual predictor variables on all 4 target variables and checking the resultant p-value.

#### Chi-Square Test for Independence

A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.

In our case, we need to determine whether there is indeed a relationship between a predictor variable and any of the target variables to a significant degree. We only need to consider these features for our further analysis.

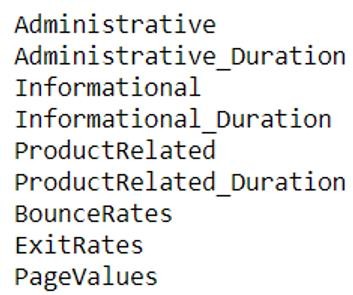
**The null hypothesis of the Chi-Square** test is that no relationship exists on the categorical variables being tested. I.e. they are independent.

The p-value will tell us if our test results are significant or not. In order to perform a chi square test and get the p-value, you need two pieces of information:

* + - Degrees of freedom. That’s just the number of categories minus 1.
    - The alpha level (α). The usual alpha or significance level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

*We reject the null hypothesis when the P-value is less than the set significance level.*

#### These are the attributes we got as significant from this chi-square method:



**9. IMPLEMENTATION**

Modelling using the listed algorithms was done under multiple scenarios and results compiled for overview.

#### To note:

For Logistic Regression, the data was first normalized with the Min-Max Scaler:

*An alternative approach to Z-score normalization; in this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.*

*A Min-Max scaling is typically done via the following equation:*



**SCENARIO DESCRIPTIONS:**

#### ORIGINAL DATA

* + - Modelling was done on the original data after default data cleaning and scaling where necessary.
    - After that we have applied statistical test in that we find some insights, since our alpha (0.05) value is greater than our p-value we will reject Ho saying that data is not normally distributed Hence our statistical test proves the same. I.e. output variable is not normally distributed.
    - Then we have applied the one-way ANOVA is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups.
    - The Independent Samples ttest compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. For this we will separate numerical features and categorical features.
    - The Chi-square test is intended to test how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.
    - Now we will make models with the significant features and note the changes if any.

#### DERIVED VARIABLES

* + - New features were engineered and added from the intuitive understanding of the attributes in the data set and were added to the existing attributes.
      * df\_new['Administrative\_Duration']: Here, we have duration in range0 is no duration,0 to 500 low duration,500 and above high
      * df\_new ['Informational\_Duration']: we can see there are 9925 with 0 duration so it’s

Better to divide into two groups with duration and without duration.

* + - * df\_new ['ProductRelated\_Duration']: Here, we have product duration in range 0 as no duration, 1 as 0 to 600, and 2 as above 60.

#### IMPORTANT FEATURES

* + - Using the “feature\_importances\_” feature of tree-based algorithms, only features that were significant to predicting the target variable Revenue were considered for modelling.

#### UNDER SAMPLING

* + - Under-Sampling decreases the number of samples of the majority Revenue on which the models are trained so that a more balanced proportion of representation of the minority Revenue is achieved.

#### OVER SAMPLING

* + - Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority Revenue in the sample.

#### SMOTE

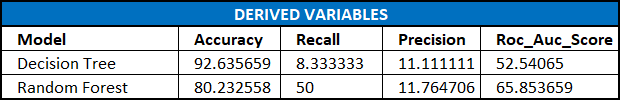
* + - Oversampling using SMOTE not only increases the size of the training data set, it also increases the variety. It creates new (artificial) training examples based on the original training examples and adds them as synthetic data points on which the model can be build.
    - [The most popular of such algorithms is called SMOTE or the Synthetic Minority Over-sampling Technique. As its name suggests, SMOTE is an oversampling method. It works by creating synthetic samples from the minor class instead of creating copies.

# 10. RESULTS AND COMPARISON STUDY

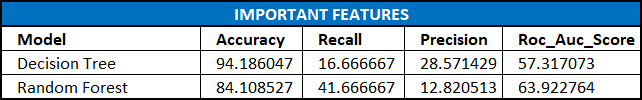
Since we are dealing with medical data, we will be focusing on **Recall** rather than Precision I.e., we will allow Type I error to creep in to our models because in our case, Type II error is far costlier.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DERIVED VARIABLE** | | | | |
| **Model** | **Accuracy** | **Recall** | **precision** | **Roc\_Auc\_score** |
| **Logistic Regression** |  |  |  |  |
| **Random Forest** |  |  |  |  |

**INFERENCE:** In this case we have high accuracy; but we neither have precision nor recall. The model is not able to identify the affected people at a reasonable rate. The accuracy is high only because of the imbalance between the non-affected and affected people. We cannot rely on these.

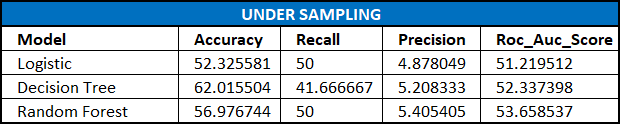


**INFERENCE:** Even after including the new derived variables, we are not able to achieve an optimal point for recall and accuracy. This case Random Forest is working the best with 50% recall and good AUC value, but Precision is still too low.

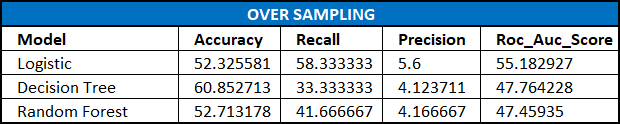


**INFERENCE:** On the derived variables, when important features were found out, these are the results we observe. It is not significantly different from the original data modelling. Even after

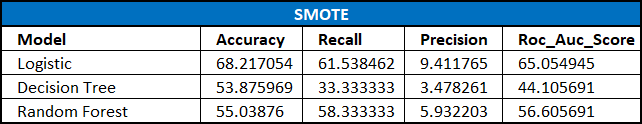
Finding imp features, we are unable to identify a satisfactory number of 1’s and the model is overfitting.



**INFERENCE:** When the data is under sampled, all the models are giving us similar results. But here we are increasing the bias error substantially as we are not considering all the data points. Even though model is working well on the recall part, the precision is almost negligible and we cannot accept this.



**INFERENCE:** Here we do not see a significant difference from when we ran the data under sampled.



**INFERENCE:** Here we get a reasonable accuracy and recall rate with decent AUC value as well with the Logistic Regression Model.

Since we had already decided to forgo Precision for our model valuation, Logistic Regression with SMOTE is the overall best model we’ve seen so far for predicting the cancer indicator with Biopsy as the target variable**.**

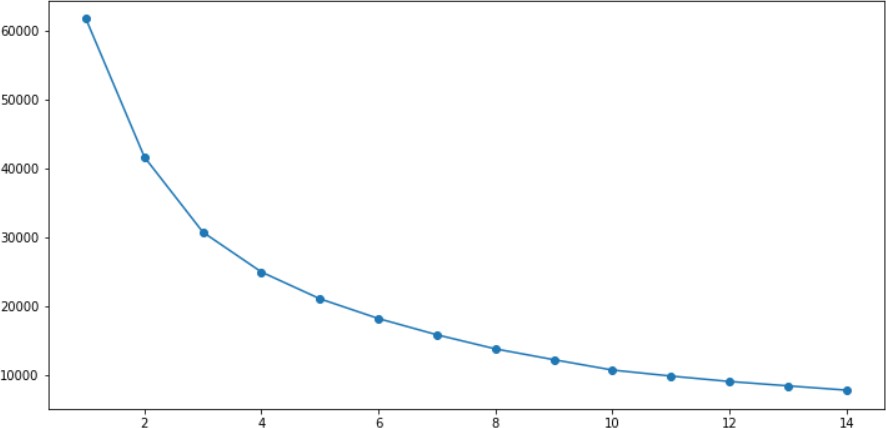
**K-Means Clustering**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, which can be used to label new data.
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

While modelling, the mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K.

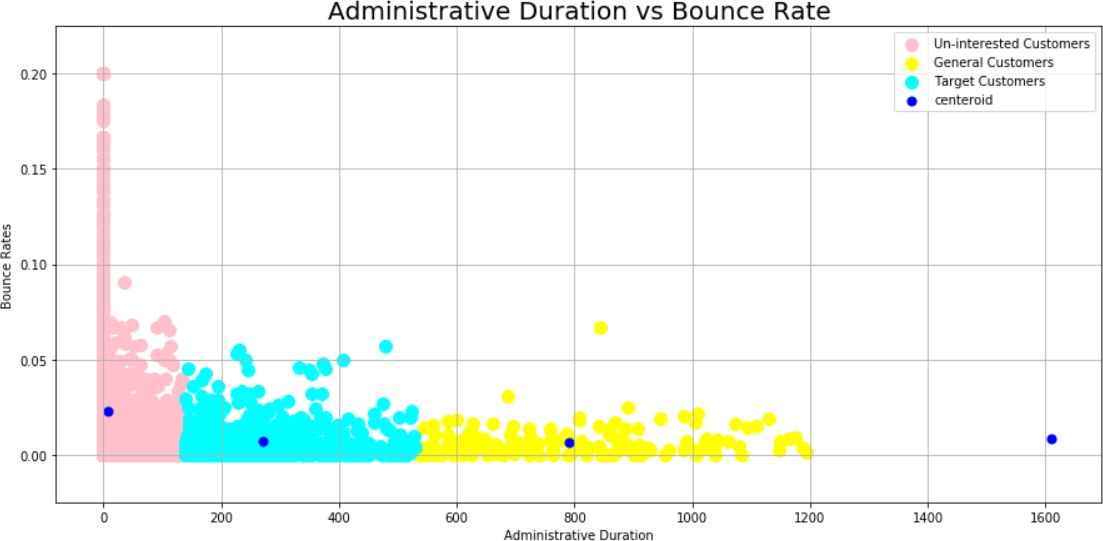


From the elbow plot, the most optimal K for our data here will be 4. We can already see a disparity here. Ideally, only two clusters should have formed: people with cancer and people without.

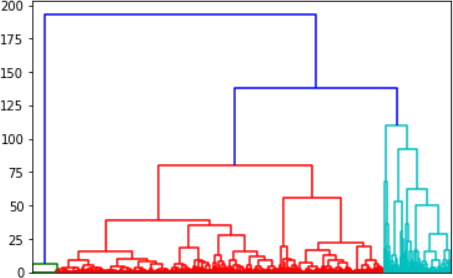
We will go ahead with the full process anyway. The result we got was:

|  |  |
| --- | --- |
| **CLUSTER** | **COUNT** |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |

### Cluster of bounce rate w.r.t administrative duration



* + Here, the plot of Agglomerative Hierarchical Clustering dendogram



We tried all methods but ward method is best here. So there can be three clusters of the customers.