

**Customer Retention**

**Submitted By**

**K. Naveen Varma**

**CUSTOMER RETENTION**

**Problem Statement:**

Customer retention is the ability of a company to retain its customers over a specified period. If a company has high customer retention, that means its customers continue to rebuy the same products, to resubscribe to the same services or, in some other way, continue buying from that company.

Customer retention is vital for two reasons: First, even if you are adding new business, you can’t grow your company if you are simultaneously losing business. The second reason is incredibly black and white: It’s always more expensive to acquire a new customer than it is to keep your current customers and sell more to them.

Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty. A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

For this dataset wants to know:

• Which variables are important to predict the variable?

• How do these variables describe which website is most friendly, feasible and reliable ?

**Business Goal:**

You are required to model the which E-Commerce Retailer is most feasible with the available independent variables. This model will then be used by the management to understand how exactly the E-Commerce Retailer vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand.

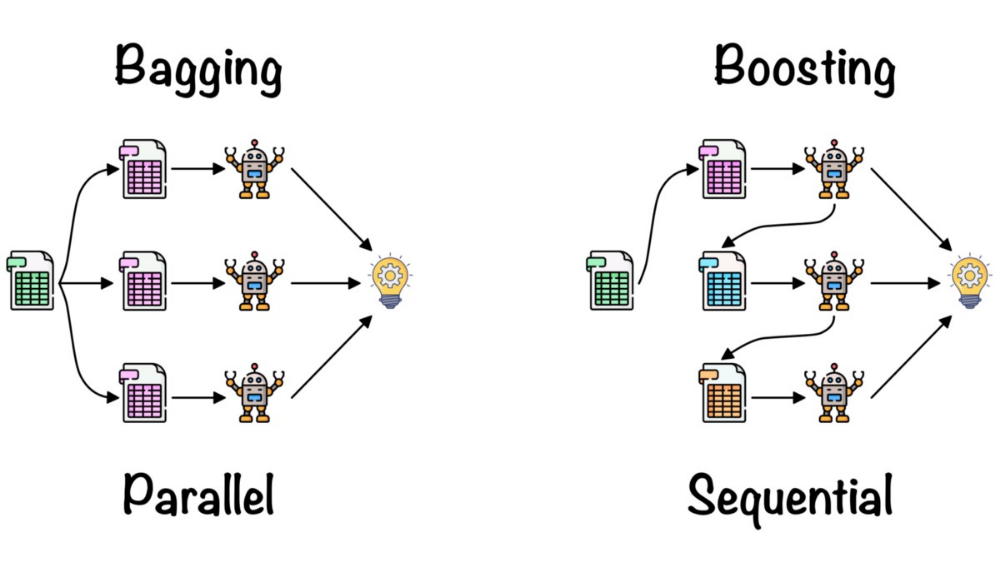
**Conceptual Background of the Domain Problem:**

Random forest is a ***Supervised Machine Learning Algorithm*** that is ***used widely in Classification and Regression problems***. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.One of the most important features of the Random Forest Algorithm is that it can handle the data set containing ***continuous variables*** as in the case of regression and ***categorical variables*** as in the case of classification. It performs better results for classification problems.

Before understanding the working of the random forest we must look into the ensemble technique. ***Ensemble***simplymeans combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

*Ensemble uses two types of methods*:1. **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example,  Random Forest.

2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example,  ADA BOOST, XG BOOST



As mentioned earlier, Random forest works on the Bagging principle. Now let’s dive in and understand bagging in detail.

#### **Bagging**

Bagging, also known as ***Bootstrap Aggregation*** is the ensemble technique used by random forest.Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as ***row sampling***. This step of row sampling with replacement is called***bootstrap***. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as ***aggregation***.

**Steps involved in random forest algorithm:**

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

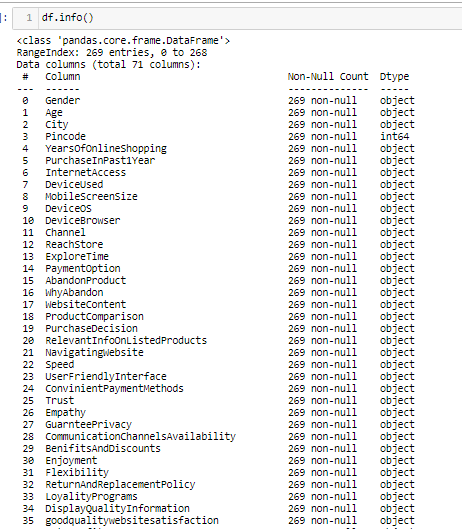
Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on ***Majority Voting or Averaging***for Classification and regression respectively.

**XGBooster:** [XGBoost](https://xgboost.ai/)is a decision-tree-based ensemble Machine Learning algorithm that uses a [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

1. A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
2. Portability: Runs smoothly on Windows, Linux, and OS X.
3. Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.
4. Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

**Data Analysis:** The Dataset Contains a Data of 269 entries each having 71 variables, in which some are numerical Data and some are Categorical Data





This is about Data set in which I changed the columns to make understandable and data set having one variable is integer remaining all are categorial values.

**Exploratory Data Analysis:**

In EDA we need to Pre-process the Data and Visualization:

Steps include in Pre-Processing Data are

1)**Data Cleaning**: - Removing Outliers, Skewness and imputing Missing Values.

2)**Data Transformation**: - like Normalization by applying normalization we can improve the accuracy and efficiency of the models. And also reduce the errors.

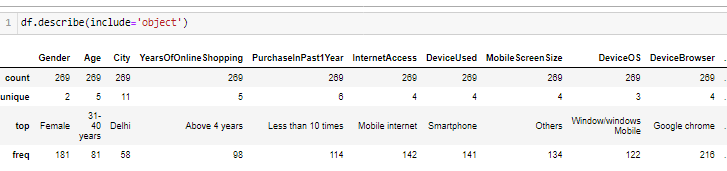
3)**Data Reduction**: By Reducing the no of features by Feature Selection Process, PCA And VIF

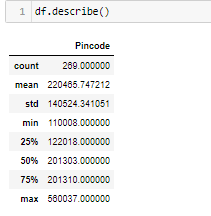
**1.Data Cleaning:** As a Part of EDA we need to do Data cleaning so firstly we need to check any null values in our data, From the below image shows we don’t have any null values, so no need to impute any data

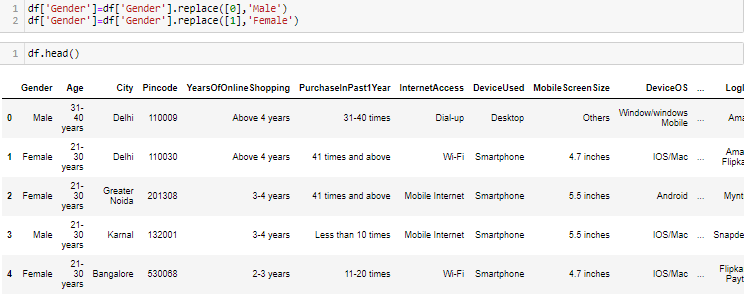
First of All Need to check the data is having any missing values



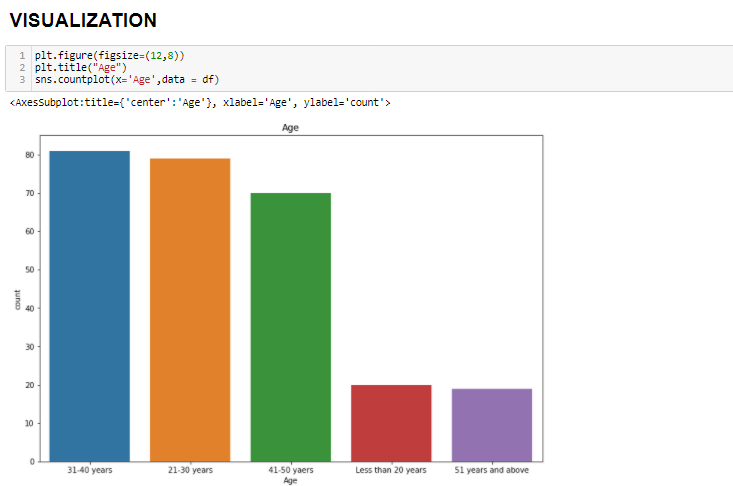




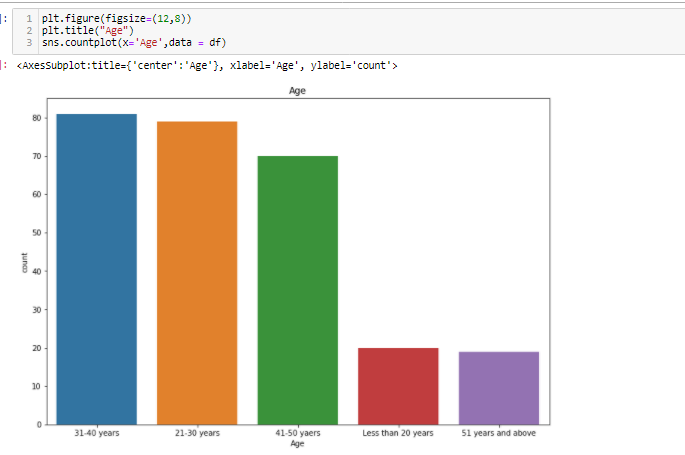




**VISUALIZATION:**

****

Here in Gender category Females are more than 175 than Male are in between 75 to 100



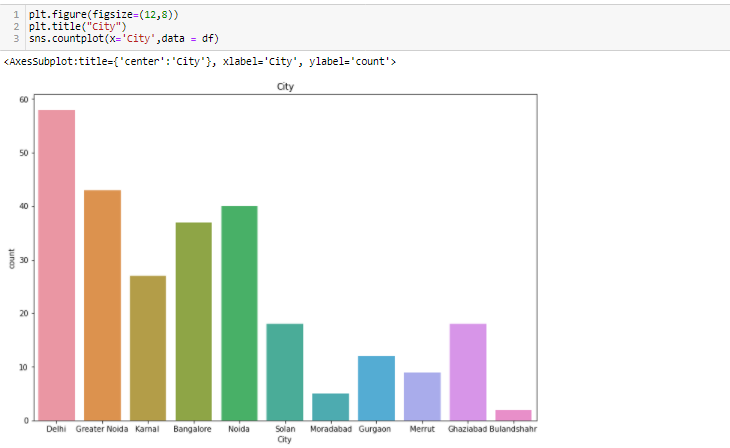
**Observations**:

From the plots we can clearly see that most of the respondents fall into the following three categories:-

1. 31-40 years
2. 21-30 years
3. 41-50 years

Very few respondents fall into the categories :-

1. Less than 20 years
2. 51 years and above.



**Observations:**

we can see a very large no.of people shopping from Delhi city i.e in between 50 to 60.

Similarly people from Greater Noida shopping in between 40 to 50.

Similarly people from Kamal shopping in between 20 to 30.

Similarly people from Banglaore shopping in between 30 to 40.

Similarly people from Noida shopping almost 40.

Similarly people from Solan city shopping in between 10 to 20.

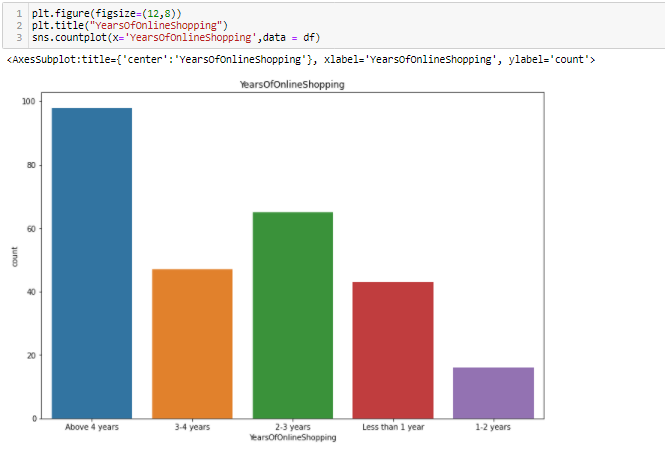
Similarly people from Moradabad shopping in between 0 to 10.

Similarly people from Gurgaon shopping in between 10 to 20.

Similarly people from Merrut shopping almost 10.

Similarly people from Ghaziabad shopping in between 10 to 20.

Similarly people from Bulandshahr shopping in between 0 to 10.



Observations:

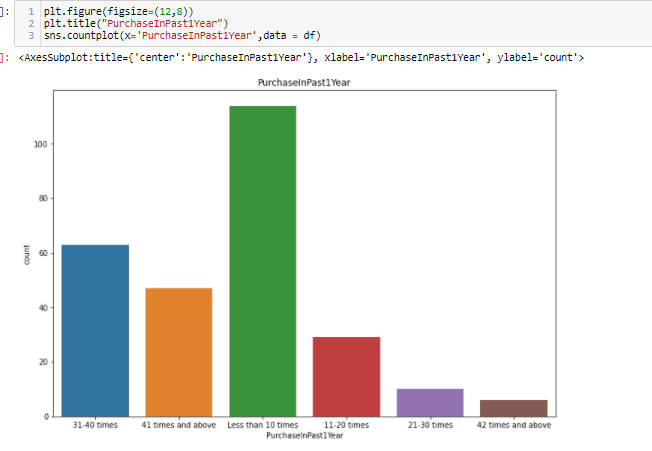
Here some E-commerce site having an experience of Above 4 years are in between 80-100.

Some other having 3-4 years in between 40-60.

Other having 2-3 years are in between 60-80.

Less than 1 year are in 40.

In between 1-2 years are in 0-20 respectively.



**Observations**:

The people in blue code above 60 in numbers are purchased the item 31-40 times in past 1 year.

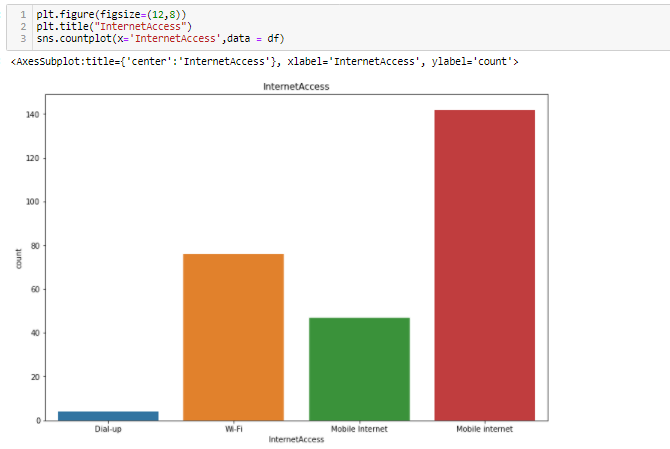
The people in orange code in between 40-60 in numbers are purchased the item 41 times in past 1 year.

The people in green code above 100 in numbers are purchased the item less than 10 times in past 1 year.

The people in red code in between 20-40 in numbers are purchased the item 11-20 times in past 1 year.

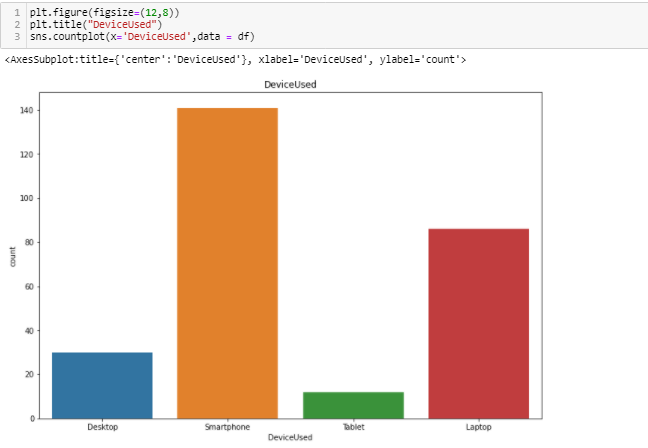
The people in violet code in between 0-20 in numbers are purchased the item 21-30 times in past 1 year.

The people in brown code in between 0-20 in numbers are purchased the item 42 times in past 1 year.



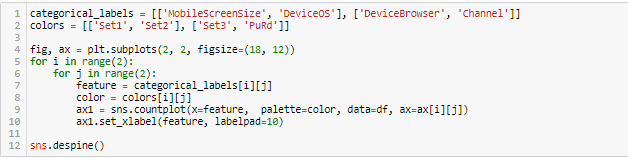
**Observations**:

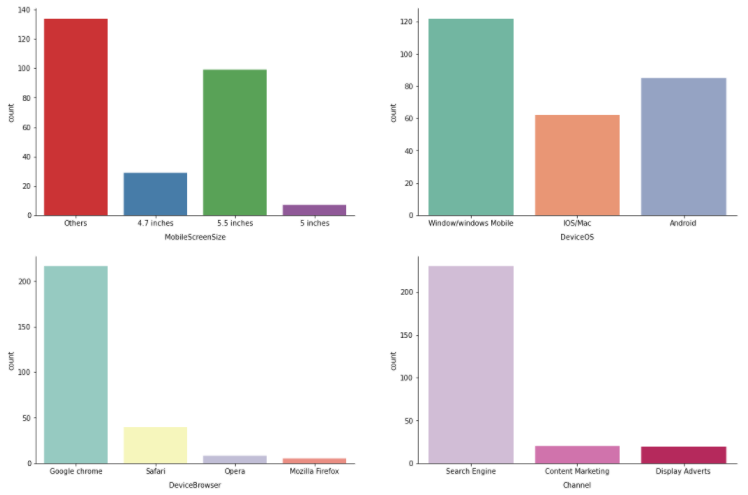
Its clear that most of people were using/Depends up on **Mobile Internet** rather than **Wifi or Dial-up** for online shopping.



Observations:

Smart phones are more popular devices used by public and then next is Laptop then Desktop & Tablet.





**Observations**:

MOBILE SCREEN SIZE:

The data which we have it is saying most people have screen size apart from 4.7,5.5,5 inches. to that most people have 5.5 inches of screensize of the mobile.

DEVICE(OS):

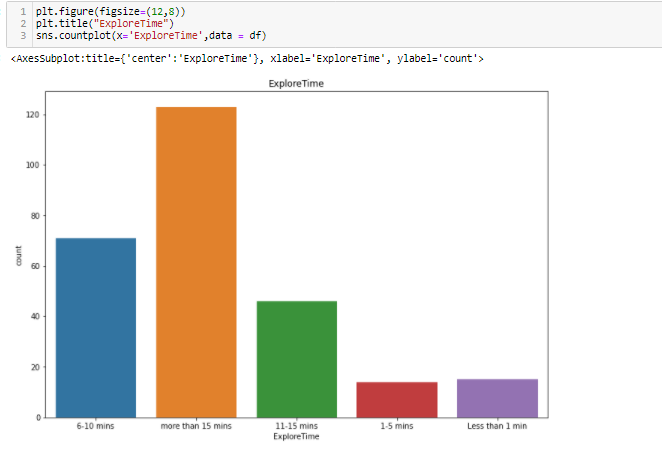
Most commonly used operating system is **Windows** and next followed by **Android** and then next **ios/Mac**.

Device Browser:

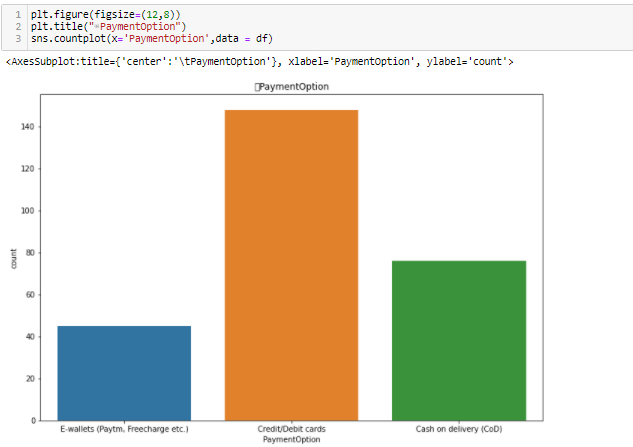
**Google Chrome** is most used browser than **safari**, **opera** and **Mozilla Firefox**.

Channel:

**Seach Engine** is the most used tool than **Content Marketing** & **Display Adverts**.

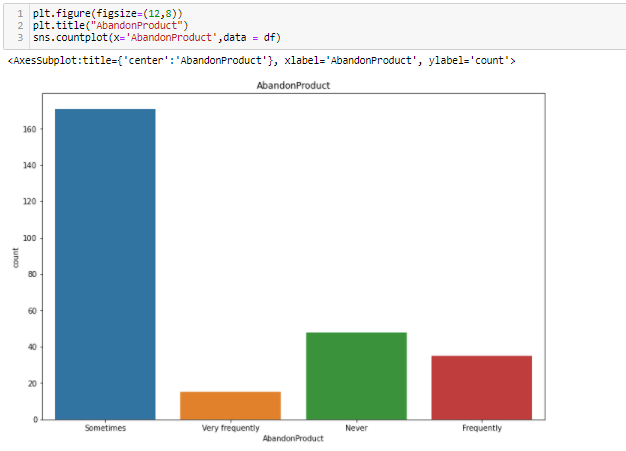


**Observation** Mostly people are engaging to explore more than 15 minutes



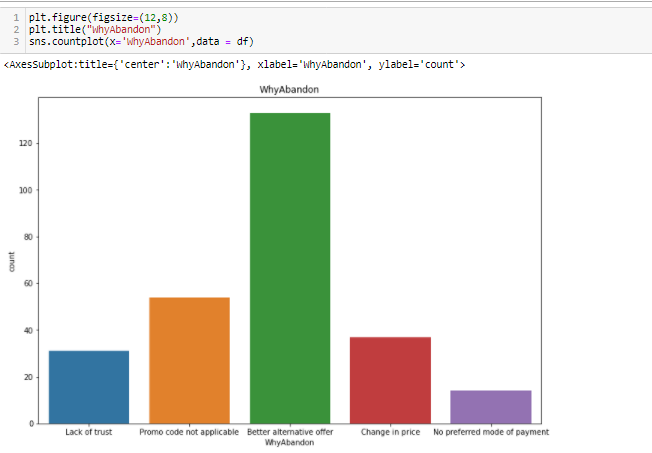
Observations:

Public used Credit/Debit Cards while shopping. Then by Cash on Deleivey(COD) and then E-wallets



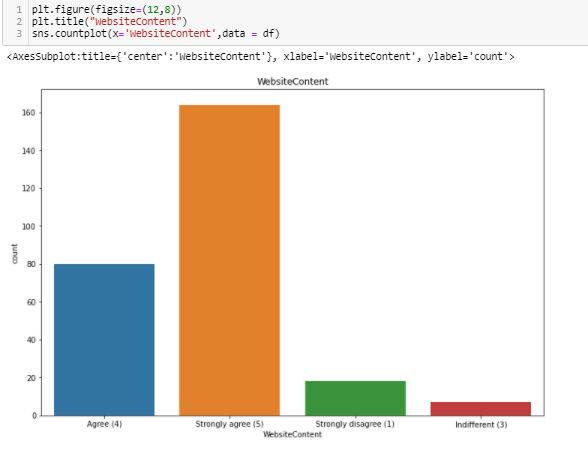
Observations:

Sometimes most of the people do abandoned the items in list while shopping online.



Observations:

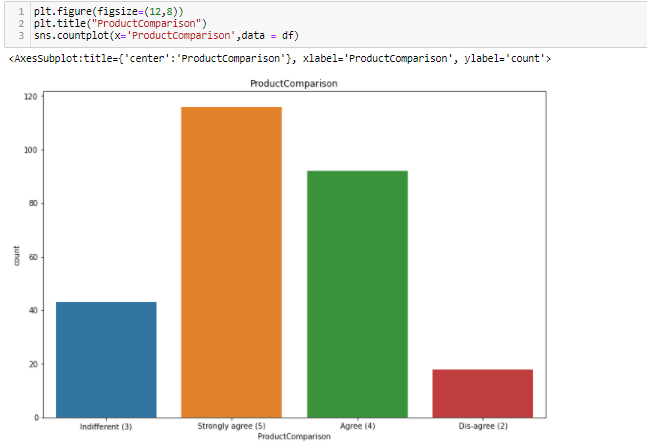
Most people abandoned the items of their list because they have " Better\_alternative\_offers" followed by "Promo code not applicable"



Observations:

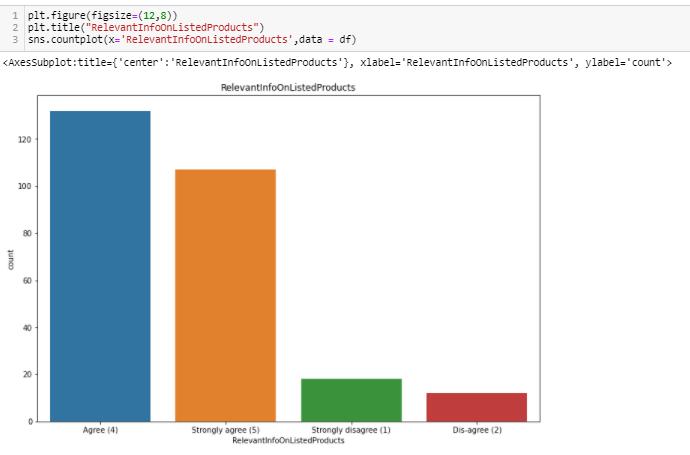
Mostly people strongly agree that "WebsiteContent" is very much important.

Very less people dis-agree with "WebsiteContent" is very much important.



Observations:

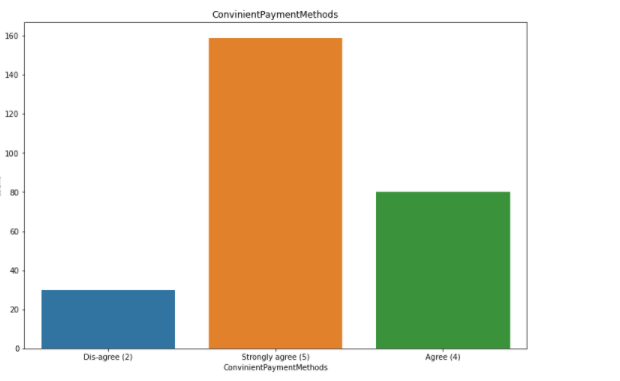
Most people thinks Product comparision is very much need. For getting the better product



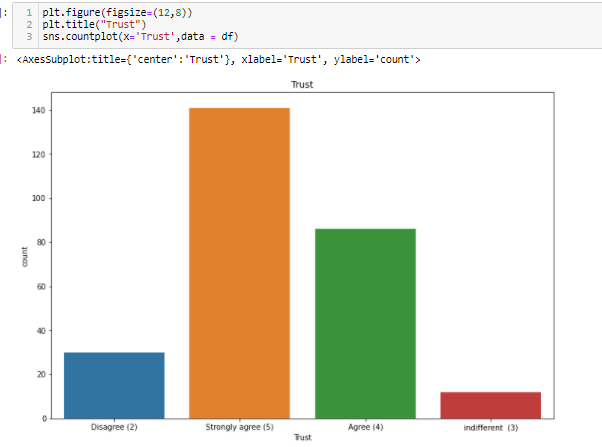
Observations

Most people will show intrest on the information about the product if they want to buy. So, there are above 120 in number of people who are agree for this.





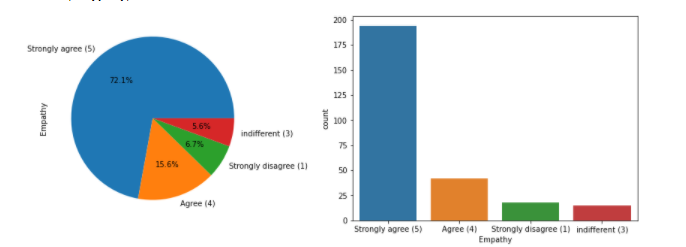
Observations:

Most people will stongly agree for purchasing the items in simple way

Observations:

Trust is the most important thing in business it plays very crucial role. So in this Obviously all people will strongly agree while purchasing an item



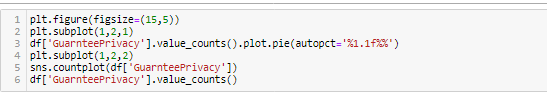


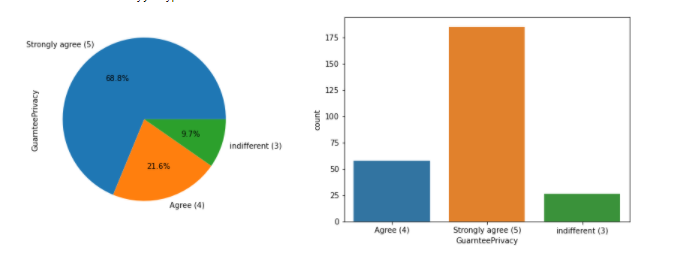
Observations:

Most of Customers firmly believe that approach of "Empathy" is the most important concern.

So here 72.1% are stongly agree to this comparing to others.

As i used pie chart here for better representation than the count plot.

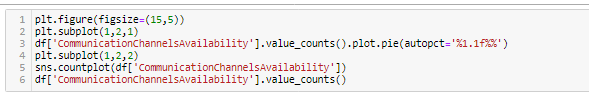


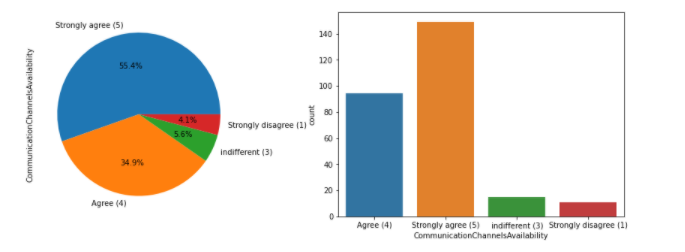


Observations:

As we all know that privacy is the most vital thing as all the customers data like credit-card number, debit-card number, are stored in the retailers database, and customer believes on the company that it will not breach the private data policy.

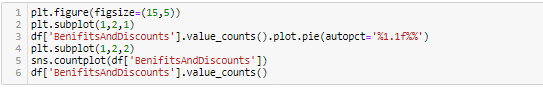
Our records also shows that most of customers strongly beleive that online store must guarantee the privacy of the data of customers

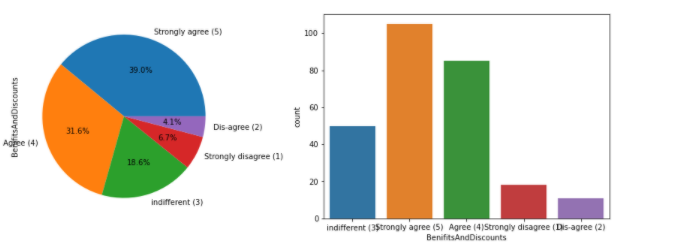




Observations:

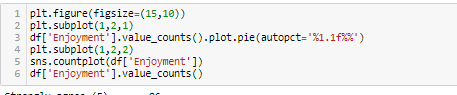
Most of the customers expect that their online store must provide the assistance to them through various modes like ("Email","Text messages","Phone Call").

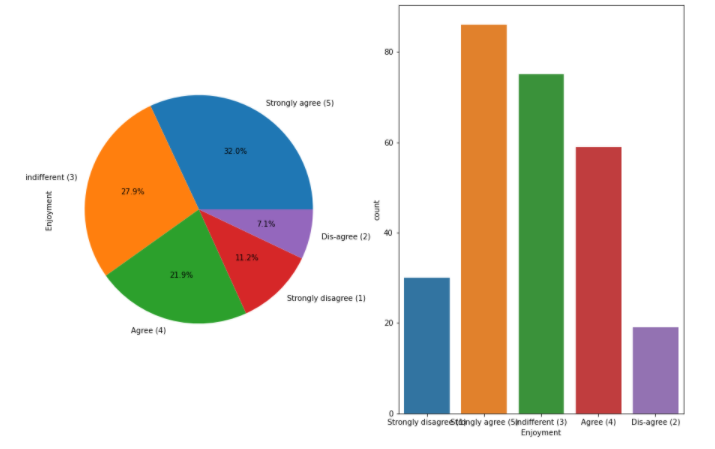




Observations:

Mostly of the customers are inclined towards the perks of monetary benefits while shopping in the forms of coupons,discount offers etc



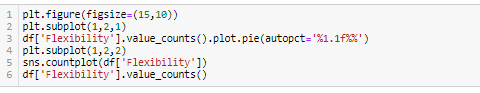


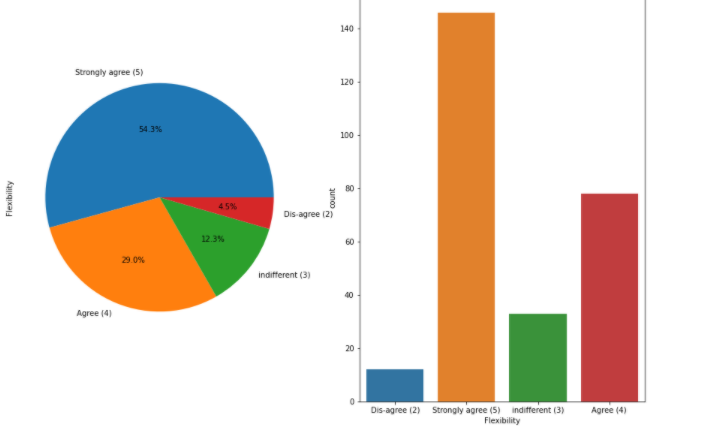
Observations:

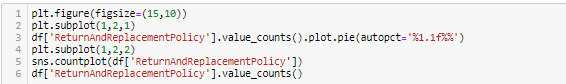
Most people agree that they enjoy the online shopping.

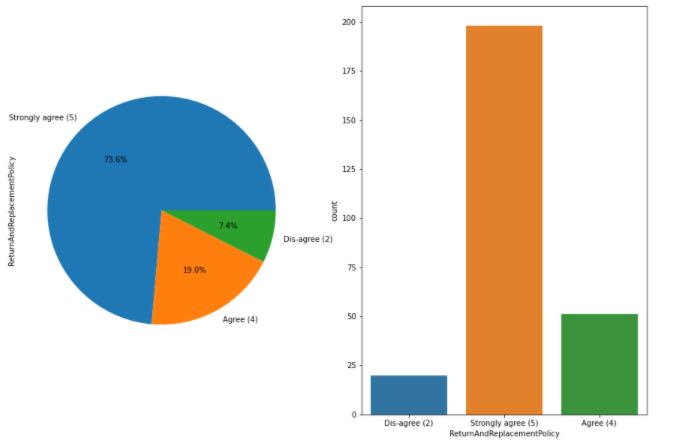
Many of respondents are also unconcerned about the enjoyment.

Very few of them disagree with the fact of enjoyment while online shopping.







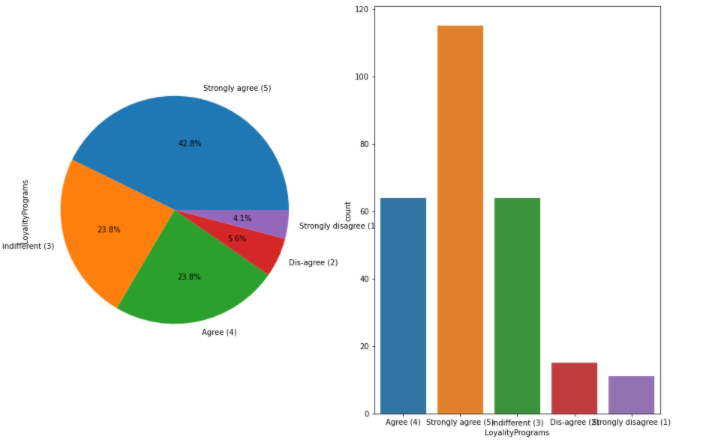


Observations:

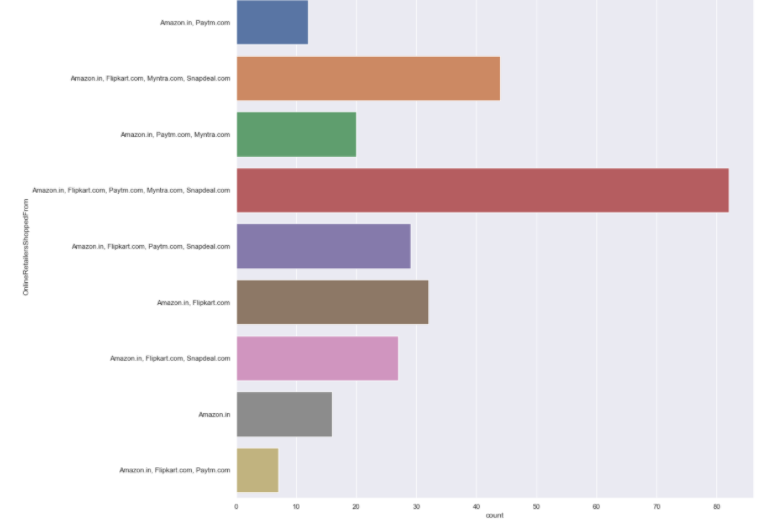
As we all know that Return and Replacement policy is very much important for a customer since if the product which he/she buyed didnt performing well then they can return it to the seller without any problem by giving proper reason of return of product.

Most of the customers are agreeing with the fact that there must be customer friendly return or replacement policy must be present in the company.





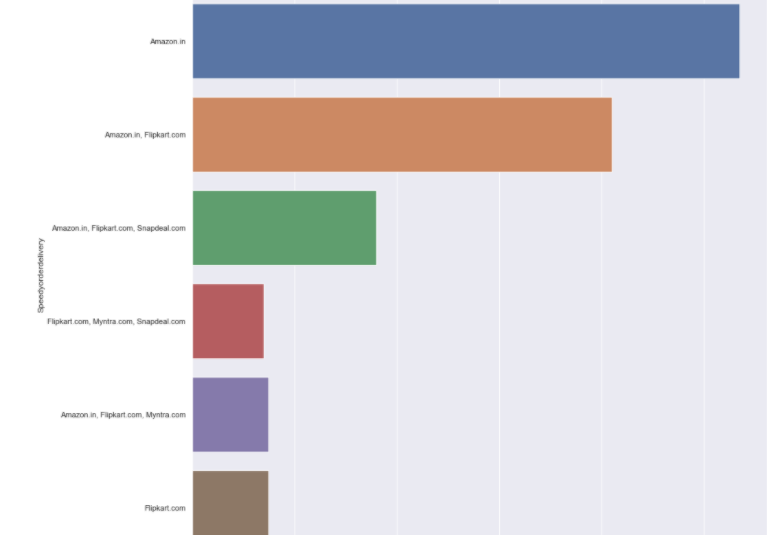




We can see that mostly people are recommending "Amazon.com" & "Flipkart.com" to others.

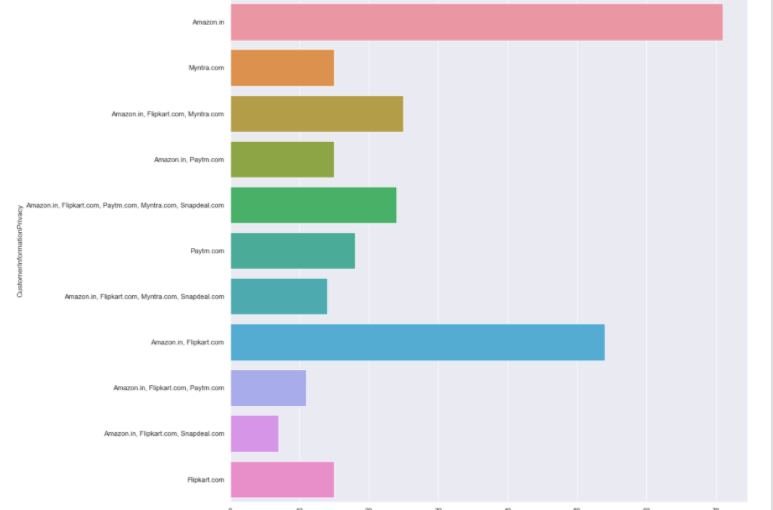
Least were recommended website is "Paytm" & "Snapdeal"





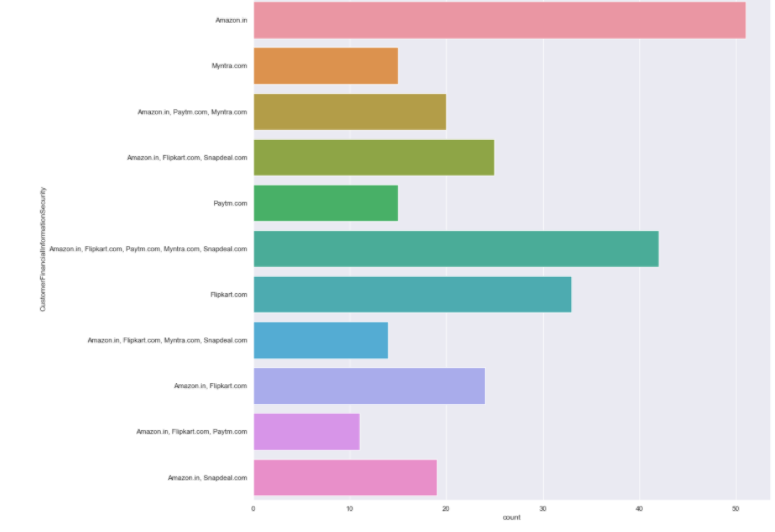
Amazon is delevering in quick way than others





Public online shopping/business privacy is good in Amazon while comparing with others.





For observing above graph analyze that Amazon leads the e-commerce industry with lot of offers and user friendly

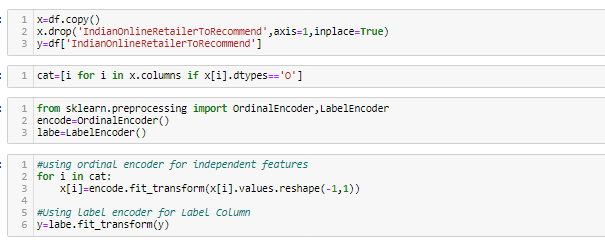
**Label Encoding**:

Label Encoding is necessary for the data to process to find any outliers are there as of our data consists of both numerical and categorical need to change categorical into numerical values using the encoding methods.

**One Hot Encoding:**

Categorical data refers to variables that are made up of **label values**, for example, a “color” variable could have the values “red“, “blue, and “green”. Think of values like **different categories** that sometimes have a natural ordering to them.

One hot encoding is one method of **converting data** to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.



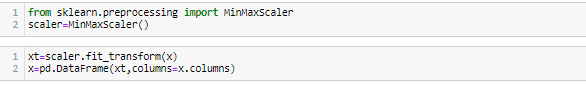
**Skewness of Data:**

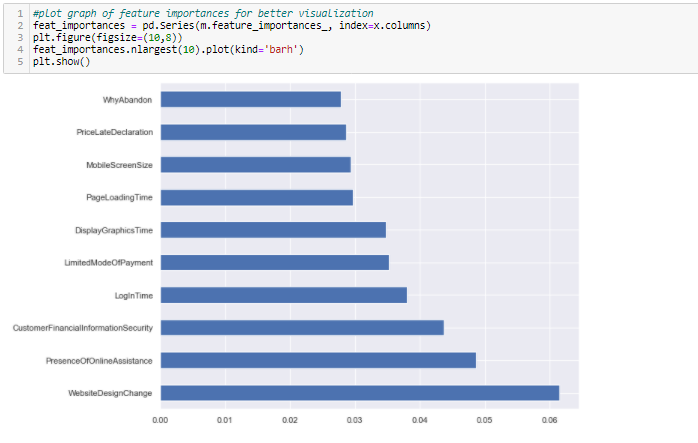
As of our numeric data is skewed we need to do normalization before go for training and testing for that need to check the skewness of data if our data is Greater than 0.5% in both positive and negative sides ,then need to do power transformation and do scaling

Scaling are of two types:

1.**Standard Scaler:** Standard scalar standardizes features of the data set by scaling to unit variance and removing the mean (optionally) using column summary statistics on the samples in the training set.

**MIN-MAX Scaler:** MinMaxScaler. For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum. MinMaxScaler preserves the shape of the original distribution.





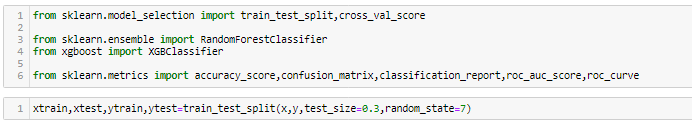
Observations:

In the above chart we can see that above features are of most importance in determining which platform will a customer recommend to his friend



**Splitting Data Into train\_test\_split: -**

This function is in sklearn. Model selection splitting the data array into two arrays. Train and Test with this function we don’t need to splitting train and test manually.by default it make random partition and we can also set the random state.it gives four o/p like x\_train, x\_test, y\_train, y\_test.



As we do splitting the data for training and testing the data then now we need to modeling

Try Different Models….

**Random Forest Classifier:** Random forest is a [supervised learning algorithm](https://builtin.com/data-science/supervised-learning-python). The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the [bagging method](https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies) is that a combination of learning models increases the overall result.

Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's .



**Now lets see about each one in our output clearly:**

**Confusion Matrix:** It is the table that is used to describe the performance of classification model on set of tests data.by using different parameters.

**Now lets understand the Recall Precision and f1-score**

**Accuracy:** it can be defined as the ratio of total number of correct classifications divided by total number of classifications.

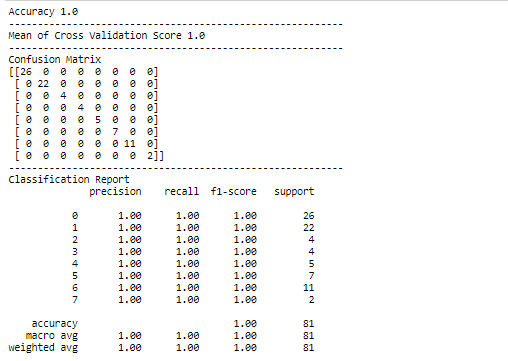
Accuracy=(TP+TN)/(TP+FP+TN+FN)

**Precision**: It is measure of all the positive predictions how many of them actually positive. Precision=TP/(TP+FP)

**F1-Score:** It give the combine result of Recall and Precision

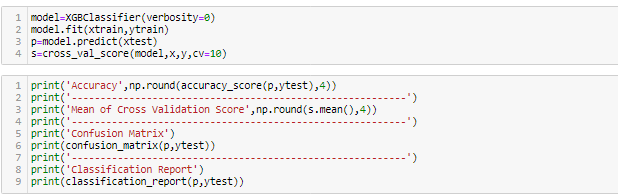
F1-score=2\*(Precision\*Recall)/ (Precision + Recall)

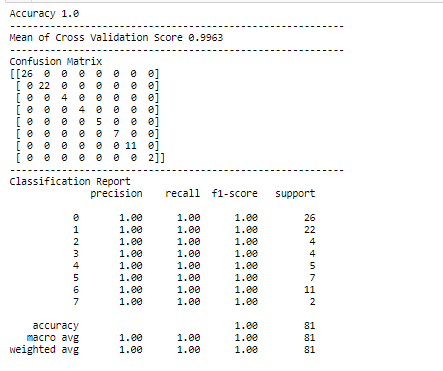
**ROC-AUC Curve:**- It is the performance measurement of the model at diff diff thresholds. ROC is the performance score and AUC is the separation score means how much mode classify 0 as 0 and 1 as 1.

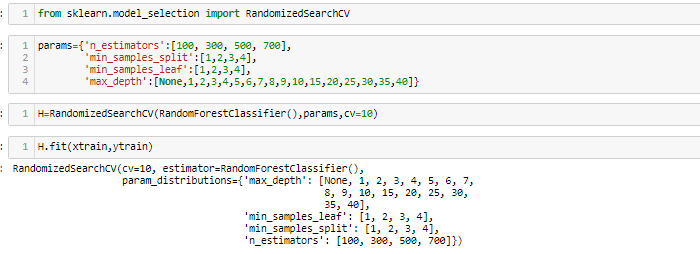


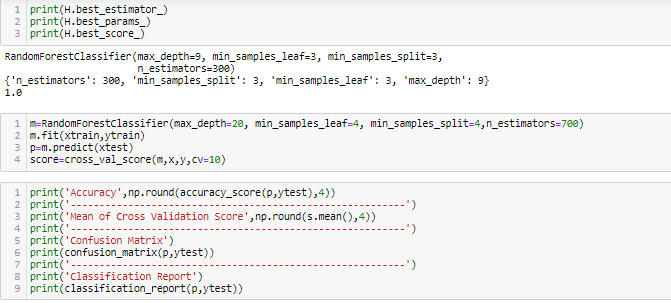
**XGBoost Classifier:**

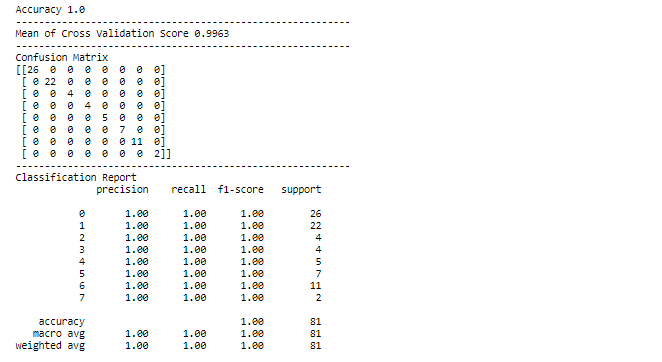
XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is **an implementation of gradient boosted decision trees designed for speed and performance**.





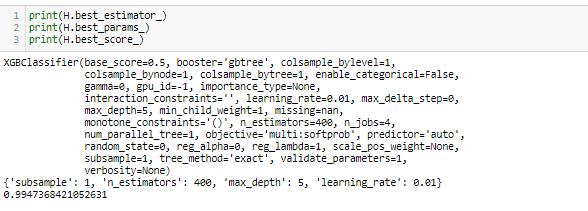


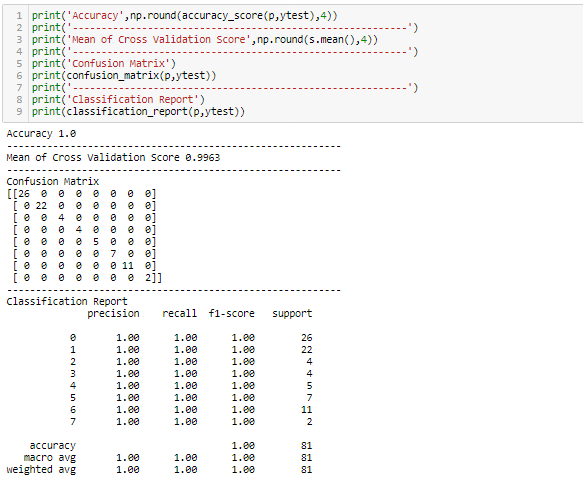




XGBoost Classifier:







**Conclusion**:

The cost of the product, the reliability of the E-commerce company and the return policies all play an equally important role in deciding the buying behaviour of online customers. The cost is an important factor as it was the basic criteria used by online retailers to attract customers. The reliability of the E-commerce company is also important, as it is even required in offline retail. It is important because customers are paying online, so they need to be sure of security of the online transaction. The return policies are important because in online retail customer does not get to feel the product. Thus, he wants to be sure that it will be possible to return the product if he does not like it in real. Whereas, the logistics factor, which included Cash on delivery option, One day delivery and the quality of packaging plays a secondary role in this process though these are Must-be-quality. This is so because these all does not interfere with the real product and people believe that this is the basic value that E-commerce websites provide.

All the websites were not equally preferred by online customers. Amazon was the most preferred followed by Flipkart. This can be explained easily by previous result that we got. These two companies are most trusted in the industry and hence, have a huge reliability. Also, the sellers listed on these websites are generally from Tier 1 cities as compared to Snapdeal and PayTM which have more sellers from tier 2 and 3 cities. Also, these websites have the most lenient return policies as compared to others and also the time required to process a return is low for these.

**Saving the model:**

****