E-retail factors for customer activation and retention

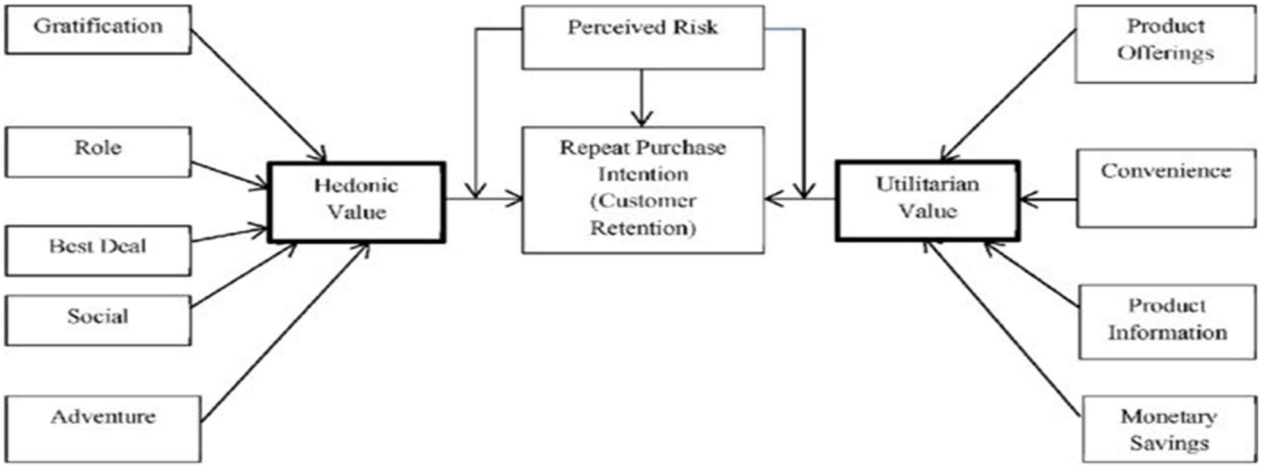
Submitted By

Amruta Shah

Objective: 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.

Problem Statement: -

To find out the e-retail success factors, which are very much critical for customer satisfaction & customer retention.



Methodology: -

The steps followed in this work, right from the dataset preparation to obtaining results are represented in Fig.

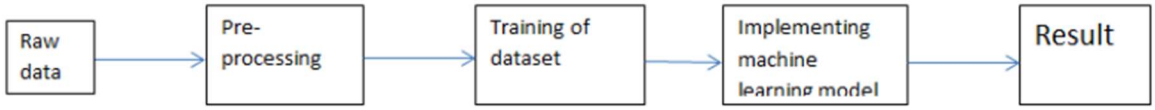


Fig1: Steps followed for obtaining results

EDA:-

We should first perform an EDA as it will connect us with the dataset at an emotional level and yes, of course, will help in building good hypothesis function.

EDA is a very crucial step. It gives us a glimpse of what our data set is all about, its uniqueness, its anomalies and finally it summarizes the main characteristics of the dataset for us.

In order to perform EDA, we will require the following python packages.

Import libraries: -

Let's use collected data set to solve the problem. For that we need to import some necessary python libraries.

```
import pandas as pd        # for data manipulation
import numpy as np         # for mathematical calculations
import seaborn as sns      # for data visualization

import matplotlib.pyplot as plt  #for graphical analysis
%matplotlib inline

from scipy.stats import zscore # to remove outliers

from sklearn.preprocessing import StandardScaler  # for normalize the model
from sklearn.preprocessing import LabelEncoder  # to convert object into int


from sklearn.model_selection import train_test_split  # for train and test model

import warnings                    # to ignore any warnings
warnings.filterwarnings("ignore")

from sklearn import metrics  # for model evaluation
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score, accuracy_score, confusion_matrix, classificat
```

Once we have imported the packages successfully, we will move on to importing our dataset.

Load Dataset: -

This collected data is in the .xlsx form so we need to use Pandas. read method to read the data.

```
data=pd.read_excel('customer_retention_dataset.xlsx',sheet_name=0)  # read the data
data
```

	1Gender of respondent	2 How old are you?	3 Which city do you shop online from?	4 What is the Pin Code of where you shop online from?	5 Since How Long You are Shopping Online ?	6 How many times you have made an online purchase in the past 1 year?	7 How do you access the internet while shopping on-line?	8 Which device do you use to access the online shopping?	9 What is the screen size of your mobile device? \t\t\t\t\t\t	10 What is the operating system (OS) of your device? \t\t\t\t	...	Longer time to get logged in (promotion, sales period)	Longer time in displaying graphics and photos (promotion, sales period)	Late declaration of price (promotion, sales period)
0	Male	31-40 years	Delhi	110009	Above 4 years	31-40 times	Dial-up	Desktop	Others	Window/windows Mobile	...	Amazon.in	Amazon.in	Flipkart.com
1	Female	21-30 years	Delhi	110030	Above 4 years	41 times and above	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	...	Amazon.in, Flipkart.com	Myntra.com	snapdeal.com
2	Female	21-30	Greater	201308	3-4 years	41 times and	Mobile	Smartphone	5.5	Android	...	Myntra.com	Myntra.com	Myntra.com

Our dataset has 269 rows and 71 columns. As we see there is too long name of the columns so, lets first rename the columns for further treatment or procedure. So, to rename the columns name let's use as below.

```
# Rename the columns of dataset for better treatment
data.columns=['gender','age','city','pin','year','shop_past_year','InternetService','deviceService','screen_size','OS','brow
```

The dataset has been successfully imported. Let's have a look at the dataset. head () gives us a glimpse of the dataset. It can be considered similar to *select * from database_table limit 5* in SQL. Let's go ahead and explore a little bit more about the different fields in the dataset. info () gives uses all the relevant information on the dataset. If your dataset has more numerical variables, consider using describe () too to summarize data along mean, median, standard variance, variance, unique values, frequency etc. isna (). sum () gives us sum of null values are present in the dataset. See below screenshot.

```
(269, 71)
gender                0
age                   0
city                  0
pin                   0
year                  0
                     ..
Longer_delivery       0
app_design            0
Frequent_disruption   0
efficient_web         0
recommendation        0
Length: 71, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 269 entries, 0 to 268
Data columns (total 71 columns):
```

#	Column	Non-Null Count	Dtype
0	gender	269 non-null	object
1	age	269 non-null	object
2	city	269 non-null	object
3	pin	269 non-null	int64
4	year	269 non-null	object
5	shop_past_year	269 non-null	object
6	InternetService	269 non-null	object
7	deviceService	269 non-null	object
8	screen_size	269 non-null	object
9	OS	269 non-null	object
10	browser	269 non-null	object
11	channel	269 non-null	object
12	how_reach	269 non-null	object
13	explore_time	269 non-null	object
14	pat_mode	269 non-null	object
15	empty_cart_time	269 non-null	object
16	abandon_bag	269 non-null	object
17	easy_content	269 non-null	object
18	pro_comp	269 non-null	object
19	comp_info	269 non-null	object
20	cleary_info	269 non-null	object
21	easy_navi	269 non-null	object
22	speed	269 non-null	object
23	user_fri	269 non-null	object
24	conv_pay	269 non-null	object
25	trust	269 non-null	object
26	empathy	269 non-null	object
27	privacy	269 non-null	object
28	communication	269 non-null	object
29	benefit	269 non-null	object
30	enjoyment	269 non-null	object
31	conv_flexi	269 non-null	object

We observe that there are 269records and 71 columns in the dataset. Dataset has object as well as numeric types. Object type in pandas is similar to strings. Now let’s try to classify these columns as Categorical, Ordinal or Numerical/Continuous.

Categorical Variables: Categorical variables are those data fields that can be divided into definite groups. In this case, Gender (Male OR Female) is categorical variables.

Ordinal Variables: Ordinal variables are the ones that can be divided into groups, but these groups have some kind of order. Like, high, medium, low. Dependents field can be considered ordinal since the data can be clearly divided into 4 categories: 0, 1, 2, 3 and there is a definite ordering also. In this case we have Strongly disagree (1), Dis-agree (2), indifferent (3), Agree (4), Strongly agree (5).

Numerical or Continuous Variables: Numerical variables are those that can take up any value within a given range. In this case 'age', 'city', 'pin', 'year', 'shop_past_year', 'InternetService', deviceService', 'screen_size', 'OS', 'browser', 'channel', 'how_reach', 'explore_time', 'pat_mode', 'empty_cart_time', 'abandon_bag', 'easy_content', 'pro_comp', 'comp_info', 'cleary_info', 'easy_navi', 'speed', 'user_fri', 'conv_pay', 'trust', 'empathy', & many.

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about subject of interest and provides insights about the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands for pre-processing of data. Dataset should therefore be explored as much as possible. Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value etc. are shown in Fig. for numerical variables of our dataset.

```
# Lets understand data at high level check the stastics of dataset
data.describe(include='all')
```

	gender	age	city	pin	year	shop_past_year	InternetService	deviceService	screen_size	OS	...	Longer_time_logIn	Longer
count	269	269	269	269.000000	269	269	269	269	269	269	...	269	
unique	2	5	11	NaN	5	6	4	4	4	3	...	10	
top	Female	31-40 years	Delhi	NaN	Above 4 years	Less than 10 times	Mobile internet	Smartphone	Others	Window/windows Mobile	...	Amazon.in	
freq	181	81	58	NaN	98	114	142	141	134	122	...	57	
mean	NaN	NaN	NaN	220465.747212	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
std	NaN	NaN	NaN	140524.341051	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
min	NaN	NaN	NaN	110008.000000	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
25%	NaN	NaN	NaN	122018.000000	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
50%	NaN	NaN	NaN	201303.000000	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
75%	NaN	NaN	NaN	201310.000000	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
max	NaN	NaN	NaN	560037.000000	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	

11 rows × 71 columns

Numerical variables of the Dataset

Pre-processing: -

Pre-processing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and Label encoding scheme during model building.

As in our dataset have duplicates or spelling mistakes in some columns so we need to treat them by using unique(), .str(). & split () method.

And encoding them by using pandas get_dummies method. As below

```
# Lets separate multipal inputs and encoding them by using getdummies method
for col in cat_col1:
    df_new=data[col].str.get_dummies(sep=',').add_prefix(col+'_')
    data=pd.concat([data,df_new],axis=1)
    data.drop(col,axis = 1,inplace= True)
```

Going ahead, we will perform univariate, bivariate and multivariate analysis one by one.

Data Visualization:-

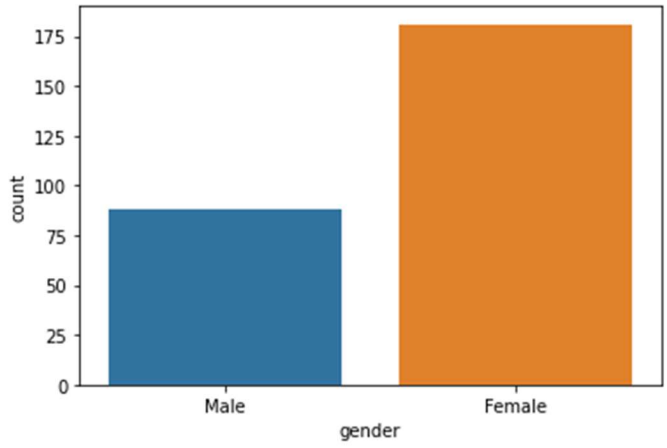
We now have a basic idea about the data. We need to extend that with some visualizations.We are going to look at three types of plots:

1. Univariate plots to better understand each variable.
2. Bivariate plots to find relationship between two variables,
3. Multivariate plots to better understand the relationships between variables.

Univariate Plots:-

We start with some univariate plots, that is, plots of each individual variable. Given that the input variables are numeric, we can create box or count plots of each. Now we are all set to perform Univariate Analysis. Univariate analysis involves analysis of one variable at a time. Let's say "Gender" then we will analyse only the "Gender" field in the dataset. The analysis is usually summarized in the form of count. For visualization, we have many options such as frequency tables, bar graphs, pie charts, histograms etc. We will stick to count charts.

```
#plot each class frequency
sns.countplot(x='gender',data=data)
plt.show()
print(data['gender'].value_counts())
```



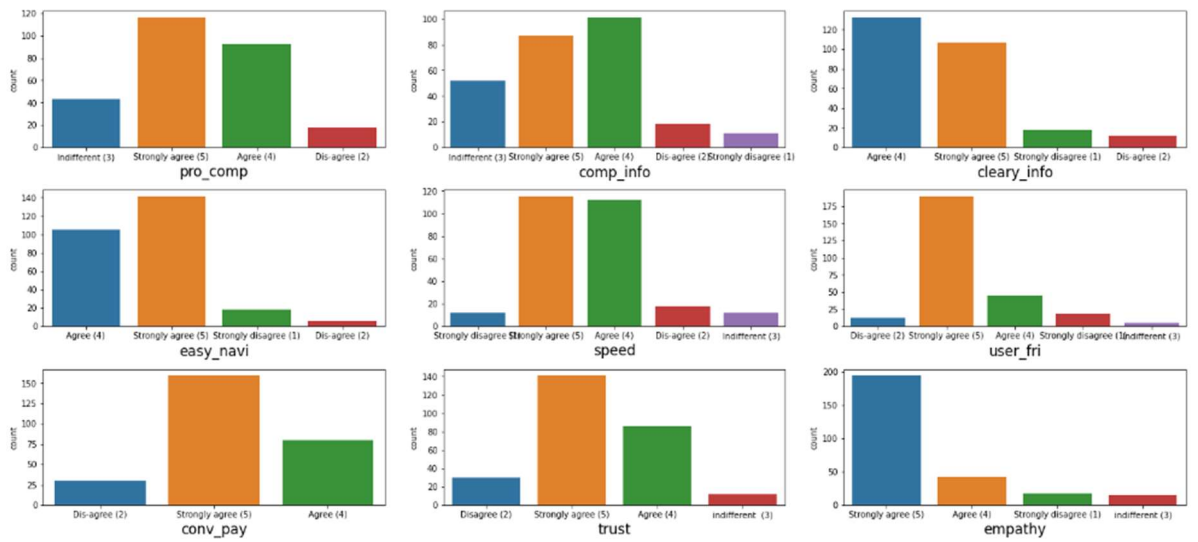

```
Female    181
Male       88
Name: gender, dtype: int64
```

From graph we can see there is female customer is more than the male customer who made online shopping.

Univariate analysis for categorical variables

Now let's move to ordinal variables.

plot() tool from pandas will help in plotting a chart of a specific kind. We can plot all the categorical variables together using plt.subplot() and give some space between them using plt.tight layout().

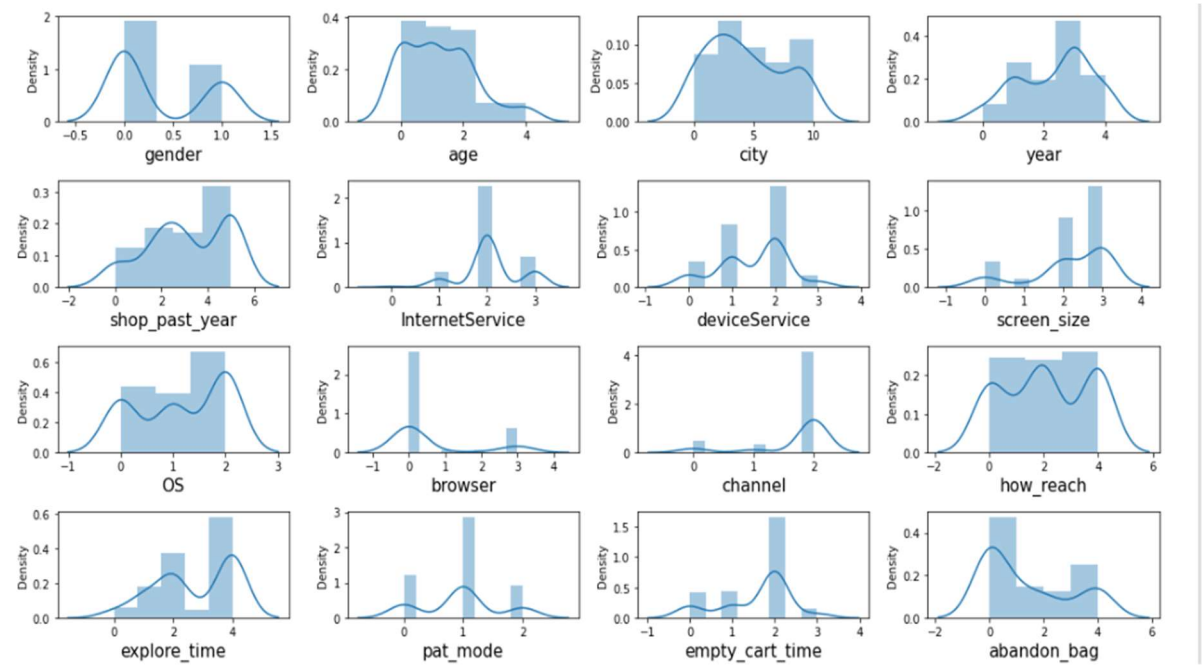


Uni-variate analysis for ordinal variables

Insights:

- There are 132 customers are given 4 star and agreed that all relevant information on listed products must be stated clearly.
- There are strongly agreed (5) 115 customers/people who thinks that the Loading and processing speed has to be there.
- From graph we can see that there are strongly agreed 189 peoples who thinks the website should be User friendly Interface.
- From graph we can see that most of people are indifferent (3) & Agree (4) for the Shopping on the website helps you fulfill certain roles.

Visualization for numerical variables will be a bit different from the ordinal and categorical variables. You may create bar plots by first creating bins, but a better plot will be a distribution, dotted line or box plot, as it will help us in identifying outliers.



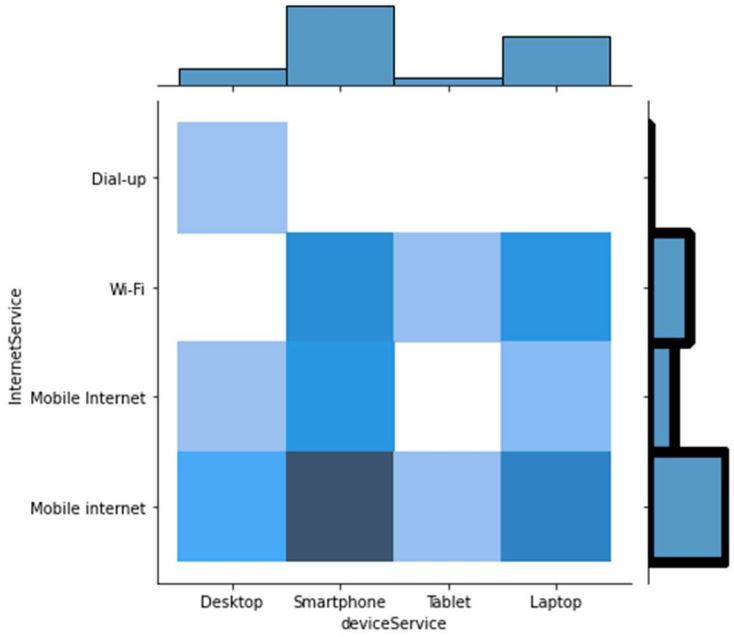
Distribution plot for continuous variables

Also, from distribution plot we can get the information about skewed data and remove it by using pre-processing techniques.
We might have to remove outliers from some columns. But that forms part of the data preparation stage.

Bivariate Plot:
Now let's find some relationship between two variables, particularly between the target variable and a predictor variable from the dataset. Formally, this is known as bivariate analysis. But as we don't have any target variable, we will check this relationship between two independent variables.
For visualization, we will be using seaborn. countplot (). It can be considered similar to the histogram for categorical variables.

```
#Bivariant graph
plt.figure(figsize =(10, 6))
sns.jointplot(x ='deviceService', y ='InternetService', data = data,kind = "hist")
plt.show()
```

```
<Figure size 720x432 with 0 Axes>
```



From graph we can see most of the peoples are used mobile internet on smartphone for online shopping.

Multivariate Plot:
Let's move on to analysing more than two variables now. We call it "Multivariate analysis". Now we can look at the interactions between the variables. First, let's look at scatterplots of all pairs of attributes. This can be helpful to spot structured relationships between input variables by using heatmap. Let's visualize the data in this correlation matrix using a heat map.

```
#check multicolinearity
myFig=plt.figure(figsize=(20,20))
sns.heatmap(data1.corr(),annot=True,annot_kws={'size':10})
plt.show()
```

Heat map matrix

Data Modelling and Observations

Correlation is used to understand the relation between a target variable and predictors. But as we don't have any target so we will get the relation between predictors only. Also, from this graph we will get the information about the multicollinearity between predictors.

Conclusion and Future Scope: -

In this paper, basics of machine learning and the associated data processing and modelling algorithms have been described, followed by their application for the task. when we compare all the variables of the data set, we find that customers who has e commerse sites which has short loading time, great deals and short loading time products.

In a nutshell….

Exploring and knowing your datasets is a very essential step. It not only helps in finding anomalies, uniqueness and pattern in the dataset but also helps us in building better hypothesis functions. If you wish to see the entire code, here is the to my jupyter notebook.