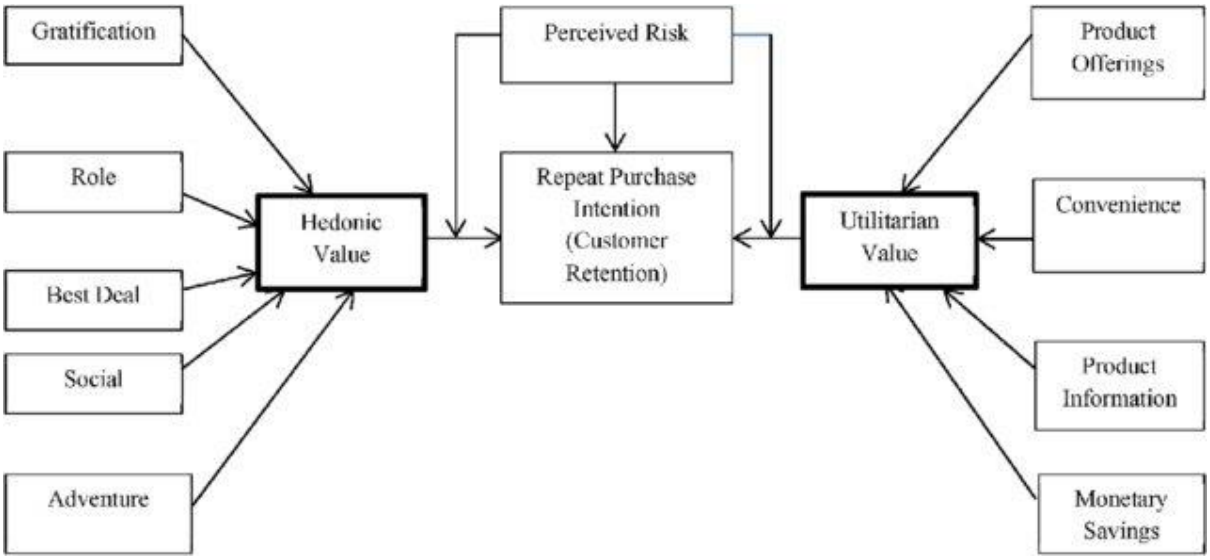
E-retail factors for customer activation and retention: A case study from Indian e-commerce customers

Customer retention refers to the activities and actions companies and organizations take to reduce the number of customer defections.

The goal of customer retention programs is to help companies retain as many customers as possible, often through customer loyalty and brand loyalty initiatives. It is important to remember that customer retention begins with the first contact a customer has with a company and continues throughout the entire lifetime of the relationship.

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 combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively.

Let us analyse the data that is being collected from the Indian online shoppers that results in the e-retail success factors, which are very much critical for customer satisfaction.

1) Loading the basic libraries required for Data Analysis

```
: #loading the libraries

#data analysis and wrangling
import pandas as pd
import numpy as np

#visualizing the data
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn import preprocessing

#for filtering the warnings
import warnings
warnings.filterwarnings("ignore")
```

2) Loading of the Dataset and checking for its shape(total rows and columns)

```
#acquiring the data
customer_retn=pd.read_excel("customer_retention_dataset.xlsx")
```

```
#checking the structure of the dataset
print(customer_retn.shape)
```

```
(269, 71)
```

3) Generating the basic information about the Dataset, i.e. column names, null value check and the datatype

```
#extracting the general information from the dataset
customer_retn.info()
```

```
269 non-null    int64
 40  41 Monetary savings
269 non-null    int64
 41  42 The Convenience of patronizing the online retailer
269 non-null    int64
 42  43 Shopping on the website gives you the sense of adventure
269 non-null    int64
 43  44 Shopping on your preferred e-tailer enhances your social status
269 non-null    int64
 44  45 You feel gratification shopping on your favorite e-tailer
269 non-null    int64
 45  46 Shopping on the website helps you fulfill certain roles
269 non-null    int64
 46  47 Getting value for money spent
269 non-null    int64
 47  From the following, tick any (or all) of the online retailers you have shopped from;
269 non-null    object
 48  Easy to use website or application
269 non-null    object
```

4) Permanently deleting the columns which are common in nature

```
#dropping the columns which are similar in kind
customer_retn.drop(customer_retn.columns[[1, 3, 6,7,8,9,10,11,20,21,22,25]], axis = 1, inplace = True)
```

5) Reviewing the first 5 data of the dataset

```
#previewing the data
customer_retn.head(5)
```

1Gender of respondent	3 Which city do 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?	13 After first visit, how do you reach the online retail store? \t\t\t\t	14 How much time do you explore the e-retail store before making a purchase decision?	15 What is your preferred payment Option? \t\t\t\t\t	16 How 4 do you abandon (selecting an items and leaving without making payment) your shopping cart? \t\t\t\t\t\t\t	17 Why did you abandon the "Bag", "Shopping Cart"? \t\t\t\t\t	18 The content on the website must be easy to read and understand	...	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	Delhi	5	4	1	3	4	3	3	4	...	Amazon.in	Amazon.in	Flipkart.com
1	Delhi	5	5	4	5	1	5	5	5	...	Amazon.in, Flipkart.com	Myntra.com	snapdeal.com
1	Greater Noida	4	5	4	4	4	3	5	5	...	Myntra.com	Myntra.com	Myntra.com
0	Karnal	4	1	1	3	1	1	2	4	...	Snapdeal.com	Myntra.com, Snapdeal.com	Myntra.com
1	Bangalore	3	2	4	5	1	4	2	5	...	Flipkart.com, Paytm.com	Paytm.com	Paytm.com

5 rows × 59 columns

6) Converting the columns' datatype from objects to integer values

```
#converting sting data to int or float data using label encoder
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

customer_retn["3 Which city do you shop online from?"] = le.fit_transform(customer_retn["3 Which city do you shop online from?"])


column= customer_retn.iloc[:, 35:]
for col in column:
    customer_retn[col] = le.fit_transform(customer_retn[col])
```

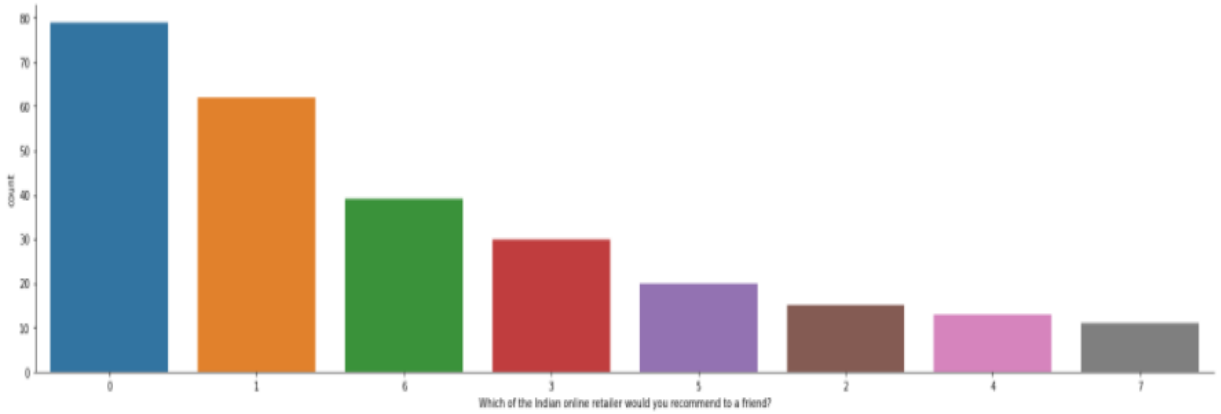
7) Visualizing the class distribution of the target value

```
#class distribution of target data
customer_retn.groupby("Which of the Indian online retailer would you recommend to a friend?").size()
```

```
Which of the Indian online retailer would you recommend to a friend?
0    79
1    62
2    15
3    30
4    13
5    20
6    39
7    11
dtype: int64
```

```
#vizual representation of the class distribution of target data
sns.catplot(x='Which of the Indian online retailer would you recommend to a friend?',kind='count',data=customer_retn,aspect=4,orc
```

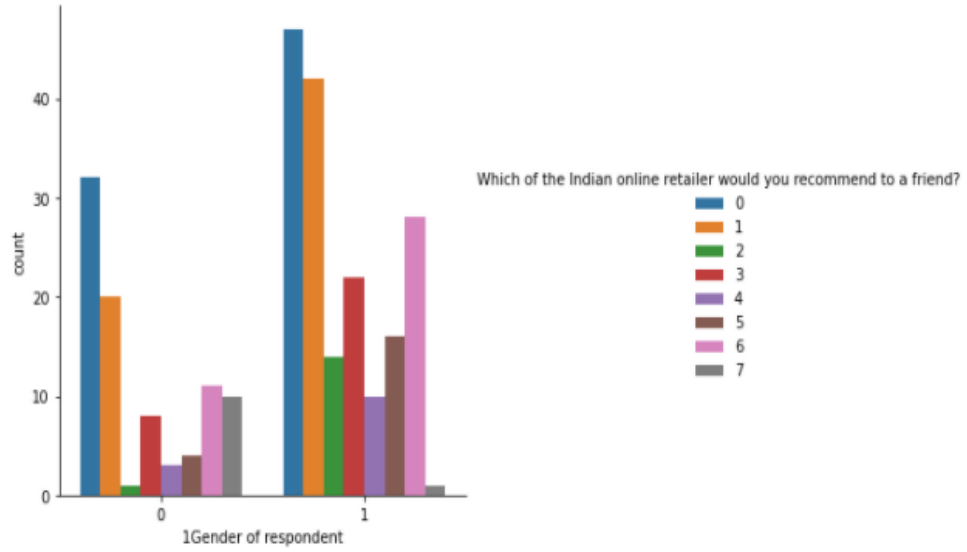
```
<seaborn.axisgrid.FacetGrid at 0x45132eb5b0>
```



8) Recommendation of an e-commerce site based upon gender reference

```
#Recommendation of an e-commerce site based upon gender(0 stands for male recommending a site and 1 stands for
#female recommending the same)
sns.catplot(x="1Gender of respondent", hue="Which of the Indian online retailer would you recommend to a friend?",
                data=customer_retn,kind='count')
```

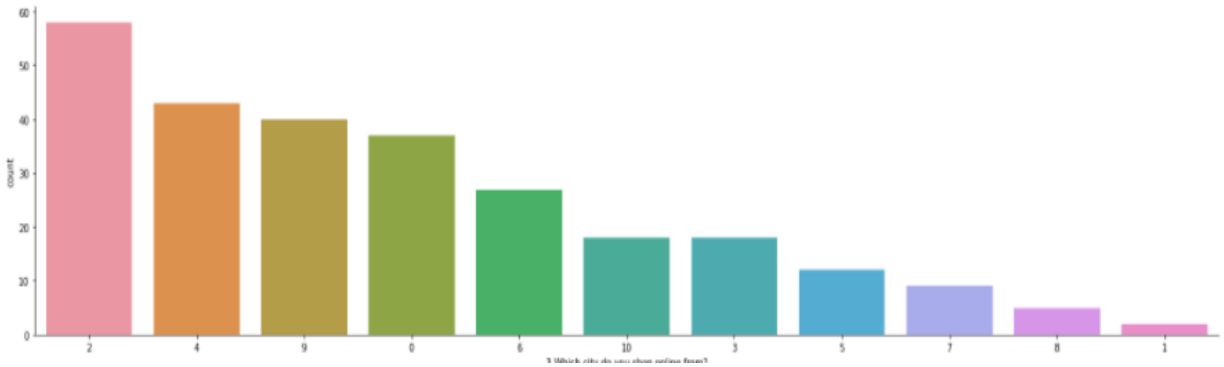
```
<seaborn.axisgrid.FacetGrid at 0x4510dd8f10>
```



9) Visual representation of cities, based upon online orders being placed

```
#Visual representation of cities, online orders are being placed from most to least
sns.catplot(x='3 Which city do you shop online from?',kind='count',data=customer_retn,aspect=4,order=customer_retn['3 Which city
```

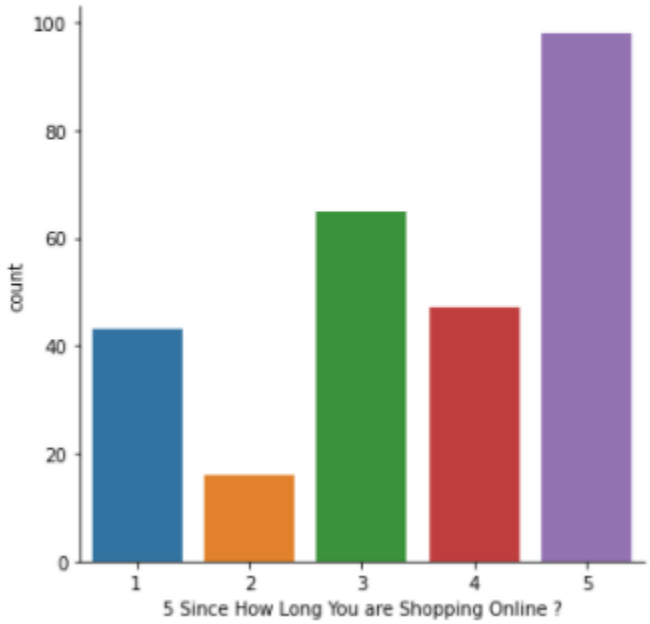
```
<seaborn.axisgrid.FacetGrid at 0x45147b0c40>
```



10) Nature of the customers being constantly shopping on online platforms(in years)

```
# To get the nature of the customers being constant on online platforms
sns.catplot(x="5 Since How Long You are Shopping Online ?",
                data=customer_retn,kind='count')
```

```
<seaborn.axisgrid.FacetGrid at 0x451477d760>
```



11) Checking for the interrelation of the columns with each other

```
#checking for the correlation(interrelation) of columns with each other
sns.heatmap(customer_retn.corr())
customer_retn.corr()
```

application													
Quickness to complete purchase	-0.060108	0.193116	0.044501	0.147161	0.364427	-0.395056	0.651272	0.120969	0.191670	0.241635	...	-0.040390	-0.05
Availability of several payment options	-0.050594	0.232500	0.116130	0.218856	0.282086	-0.272374	0.599390	0.263573	0.354669	0.353680	...	-0.031597	-0.06
Speedy order delivery	-0.085661	0.190750	-0.051109	-0.154673	0.019188	-0.547421	-0.272305	-0.311849	0.038192	0.078555	...	0.038311	0.05
Privacy of customers' information	-0.065302	-0.250461	0.039544	0.146943	0.345358	-0.235768	-0.281579	0.400294	-0.184969	-0.433154	...	-0.335177	-0.13
Security of customer financial information	0.015757	-0.262662	0.087524	0.247703	0.482184	-0.094852	-0.223870	0.308785	-0.065654	-0.169233	...	0.118340	0.28
Perceived Trustworthiness	-0.160663	-0.089781	0.088715	0.217893	0.374527	-0.239062	0.311483	0.182375	0.360375	0.188254	...	-0.001430	0.17
Presence of online assistance through	0.066122	0.019085	0.052123	0.188676	0.210917	-0.039242	0.557380	0.232494	0.236292	0.372317		0.261838	0.12

12) Analysing the statistical report of the Dataset to analyse if there is any outliers present or not by checking the mean, standard deviation and the min and max values, etc.

```
#checking for the statastical report
customer_retn.describe()
```

	1Gender of respondent	3 Which city do 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?	13 After first visit, how do you reach the online retail store? \t\t\t\t	14 How much time do you explore the e-retail store before making a purchase decision?	15 What is your preferred payment Option? \t\t\t\t\t	16 How 4 do you abandon (selecting an items and leaving without making payment) your shopping cart? \t\t\t\t\t\t\t	17 Why did you abandon the "Bag", "Shopping Cart"? \t\t\t\t\t	18 The content on the website must be easy to read and understand	...	Longer time to get logged in (promotion, sales period)	Longer time in displaying graphics and photos (promotion, sales period)
count	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	...	269.000000	269.000000
mean	0.669145	4.494424	3.524164	2.672862	2.546468	3.921933	1.784387	2.884758	2.684015	4.382900	...	4.044610	4.063197
std	0.471398	3.187687	1.436586	1.651788	1.264718	1.196014	1.084997	1.028380	1.344060	1.046603	...	3.343218	3.177536
min	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	0.000000	0.000000
25%	0.000000	2.000000	3.000000	1.000000	1.000000	3.000000	1.000000	3.000000	2.000000	4.000000	...	1.000000	1.000000
50%	1.000000	4.000000	4.000000	2.000000	3.000000	4.000000	1.000000	3.000000	2.000000	5.000000	...	3.000000	4.000000
75%	1.000000	7.000000	5.000000	4.000000	4.000000	5.000000	2.000000	3.000000	4.000000	5.000000	...	7.000000	7.000000
max	1.000000	10.000000	5.000000	5.000000	4.000000	5.000000	4.000000	5.000000	5.000000	5.000000	...	9.000000	9.000000

13) Removing the outliers

```
#removing outliers
z_score=np.abs(zscore(customer_retn))
print(customer_retn.shape)
customer_retn_final=customer_retn.loc[(z_score<3).all(axis=1)]
print(customer_retn_final.shape)
```

```
(269, 59)
(215, 59)
```

14) Separating the independent and dependent variables

```
: #separating the independent and dependent variables
  x=customer_retn.iloc[:, :-1]
  y =customer_retn.iloc[:,-1:]
```

15) Using PCA is to represent a multivariate data table as smaller set of variables

```
#The most important use of PCA is to represent a multivariate data table as smaller set of variables
from sklearn.decomposition import PCA
pca = PCA(n_components=6)
pca.fit(x)
pca_samples = pca.transform(x)
```

```
ps = pd.DataFrame(pca_samples)
ps.head()
```

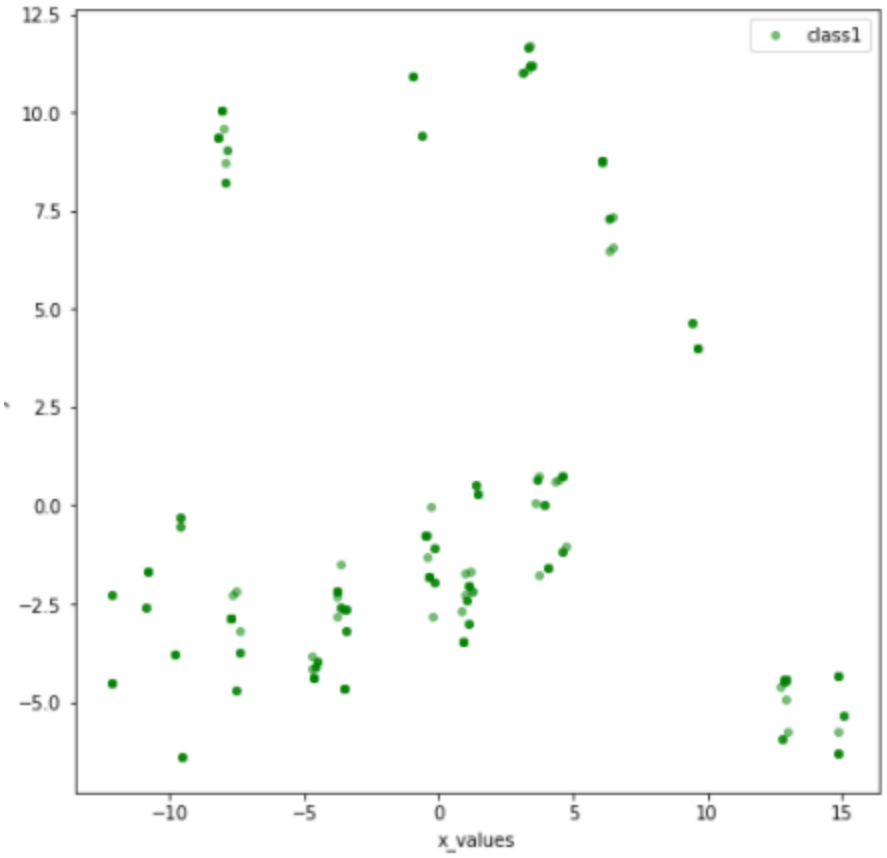
	0	1	2	3	4	5
0	14.883751	9.029534	-4.323335	-0.671339	-4.160321	5.752302
1	3.428747	-6.121247	11.193267	-4.213890	0.637243	-0.206265
2	12.923859	-1.642095	-4.410792	3.605532	0.346543	-3.661024
3	-0.361631	-3.221160	-1.813689	2.189595	4.214044	5.960117
4	6.107019	-8.985150	8.792701	-2.881602	-2.684286	-0.884366

```
#I have chosen the (PC3,PC1) pair. Since each component is the projection of all the points of the original dataset
from mpl_toolkits.mplot3d import Axes3D
from mpl_toolkits.mplot3d import proj3d
tocluster = pd.DataFrame(ps[[0,2]])
print (tocluster.shape)
print (tocluster.head())

fig = plt.figure(figsize=(8,8))
plt.plot(tocluster[0], tocluster[2], 'o', markersize=4, color='green', alpha=0.5, label='class1')

plt.xlabel('x_values')
plt.ylabel('y')
plt.legend()
plt.show()
```

```
(269, 2)
          0         2
0  14.883751 -4.323335
1   3.428747 11.193267
2  12.923859 -4.410792
3  -0.361631 -1.813689
4   6.107019  8.792701
```



16) Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

clusterer = KMeans(n_clusters=4,random_state=42).fit(tocluster)
centers = clusterer.cluster_centers_
c_preds = clusterer.predict(tocluster)
print(centers)
```

```
[[ 2.28945165 -0.77808413]
 [-0.13707483  9.65372001]
 [13.63483286 -4.9738333 ]
 [-6.80147024 -3.2531446 ]]
```

17) Predicting 100 results

```
print (c_preds[0:100])
```

```
[2 1 2 0 1 0 0 1 0 3 0 3 3 3 2 0 1 0 0 0 1 1 0 3 0 3 3 3 2 0 1 2 1 2 0 1 0
 0 1 0 3 0 3 3 2 1 2 0 0 1 0 0 1 1 0 3 0 3 3 2 1 2 0 0 1 1 2 1 2 0 1 3 3 3
 3 3 0 1 2 1 0 0 3 0 3 0 1 0 3 0 3 3 3 3 2 1 0 0 3 3]
```

18) Let's check out what are the top 10 features people rely upon of each cluster.

```
c0.sort_values(ascending=False)[0:10]
```

```
Privacy of customers' information                                                                  7.630435
Security of customer financial information                                                         6.260870
Fast loading website speed of website and application                                              5.543478
41 Monetary savings                                                                                5.000000
29 Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)    4.847826
32 Shopping online is convenient and flexible                                                      4.847826
33 Return and replacement policy of the e-tailer is important for purchase decision                4.847826
35 Displaying quality Information on the website improves satisfaction of customers                4.847826
36 User derive satisfaction while shopping on a good quality website or application                4.847826
40 Provision of complete and relevant product information                                          4.847826
dtype: float64
```

```
c1.sort_values(ascending=False)[0:10]
```

```
Longer page loading time (promotion, sales period)
8.602410
Limited mode of payment on most products (promotion, sales period)
6.819277
Longer time in displaying graphics and photos (promotion, sales period)
6.445783
Frequent disruption when moving from one page to another
5.771084
Late declaration of price (promotion, sales period)
5.506024
33 Return and replacement policy of the e-tailer is important for purchase decision
5.000000
24 User friendly Interface of the website
4.903614
37 Net Benefit derived from shopping online can lead to users satisfaction
4.903614
14 How much time do you explore the e- retail store before making a purchase decision?
4.867470
27 Empathy (readiness to assist with queries) towards the customers
4.807229
dtype: float64
```

```
c2.sort_values(ascending=False)[0:10]
```

```
Longer time to get logged in (promotion, sales period)
8.078125
Longer page loading time (promotion, sales period)
7.390625
Longer time in displaying graphics and photos (promotion, sales period)
6.093750
Fast loading website speed of website and application
6.015625
Security of customer financial information
5.906250
From the following, tick any (or all) of the online retailers you have shopped from;
5.843750
Presence of online assistance through multi-channel
5.453125
41 Monetary savings
5.000000
24 User friendly Interface of the website
5.000000
28 Being able to guarantee the privacy of the customer
5.000000
dtype: float64
```

```
c3.sort_values(ascending=False)[0:10]
```

```
3 Which city do you shop online from?                                                                  6.118421
Availability of several payment options                                                                4.552632
28 Being able to guarantee the privacy of the customer                                                 4.526316
27 Empathy (readiness to assist with queries) towards the customers                                    4.355263
18 The content on the website must be easy to read and understand                                      4.342105
38 User satisfaction cannot exist without trust                                                        4.342105
36 User derive satisfaction while shopping on a good quality website or application                    4.328947
29 Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)  4.289474
20 Complete information on listed seller and product being offered is important for purchase decision.  4.223684
Longer page loading time (promotion, sales period)                                                     4.197368
dtype: float64
```

In the same way we can smaller the set of independent variables by selecting top 40-50 features and training the required models to decide upon the retention factor of the customers.