A

Data science

Project

on

“Customer Retentation”

Submitted by:

Santosh Arvind Dharam

ACKNOWLEDGMENT

I feel great pleasure to present the Project entitled “customer retentation”. But it would be unfair on our part if I do not acknowledge efforts of some of the people without the support of whom, this Project would not have been a success. First and for most I am very much thankful to my respected SME ‘shrishti Maan’ for his leading guidance in this Project. Also he has been persistent source of inspiration to me. I would like to express my sincere thanks and appreciation to ‘flip robo’ for their valuable support. Most importantly I would like to express our sincere gratitude towards my Friend & Family for always being there when I needed them most.

Mr. Santosh Arvind Dharam

INTRODUCTION

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

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

Data Scientists have to apply their analytical skills to give findings and conclusions in detailed data analysis

Analytical Problem Framing

EDA steps:

1) import necessary libraries:

first we will import all the necessary libraries which will be usefull for analysis of data

```
In [1]: #import all libraries
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        warnings.filterwarnings('ignore')
        from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
        from sklearn.linear_model import LogisticRegression,Lasso,LinearRegression
        from sklearn.neighbors import KNeighborsRegressor
        from sklearn.svm import SVR
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.ensemble import AdaBoostRegressor,GradientBoostingRegressor
        from sklearn.preprocessing import LabelEncoder,StandardScaler
        from sklearn.model_selection import train_test_split,GridSearchCV
        from sklearn.decomposition import PCA
        from scipy.stats import zscore
        from sklearn.model_selection import cross_val_score
```

in thid case we have to import all the necessary library that are usefull for data analysis in jupyter notebook

2)extract the dataset in jupyter notebook:

now lets we will extract the data in our jupyter notebook by using pandas library

```
2]: data=pd.read_excel("C:\\Users\\SAI BABA\\Desktop\\customer_retention_dataset.xlsx")
    data.head()
```

2]:

	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	10 What is the operating system (OS) of your device\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)	Long loadi (pro sales
0	0	3	Delhi	110009	5	4	4	3	5	1	...	Amazon.in	Amazon.in	Flipkart.com	Flipk
1	1	2	Delhi	110030	5	5	2	1	2	3	...	Amazon.in, Flipkart.com	Myntra.com	snapdeal.com	Snapd
2	1	2	Greater Noida	201308	4	5	3	1	4	2	...	Myntra.com	Myntra.com	Myntra.com	Myn
3	0	2	Karnal	132001	4	1	3	1	4	3	...	Snapdeal.com	Myntra.com, Snapdeal.com	Myntra.com	Pay
4	1	2	Bangalore	530068	3	2	2	1	2	3	...	Flipkart.com, Paytm.com	Paytm.com	Paytm.com	Pay

5 rows × 71 columns

table shows the data along various columns in integer and float type ,there are total 71 columns

Data is extracted for further analysis in jupyter notebook

2) Encoding the dataset:

In this case as our data contains some categorical column having object type of data it is necessary to convert it into numerical form by LabelEncoder

```
data["Easy to use website or application"]=label
data
```

Out[28]:

	1Gender of respondent	2 How old are you	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	10 What is the operating system (OS) of your device\t\t	11 What browser do you run on your device to access the website	...	Speedy order delivery	Availability of several payment options	Quickness to complete purchase	Reliability of the website or application	ap
0	0	3	110009	5	4	4	3	5	1	1	...	0	9	8	9	
1	1	2	110030	5	5	2	1	2	3	1	...	1	2	2	8	
2	1	2	201308	4	5	3	1	4	2	1	...	0	10	5	6	
3	0	2	132001	4	1	3	1	4	3	2	...	3	2	3	3	
4	1	2	530068	3	2	2	1	2	3	2	...	0	4	4	6	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
264	1	2	173212	2	1	3	1	4	2	4	...	0	0	0	0	
265	1	3	201008	2	4	3	1	5	2	1	...	4	7	6	7	
266	1	4	560010	3	1	3	2	5	1	1	...	0	1	0	0	
267	1	1	173229	3	1	2	1	4	2	1	...	5	8	7	2	
268	1	4	201009	3	4	3	1	4	2	1	...	0	0	0	0	

269 rows × 71 columns

so table contains all 71 column and they are having integer values in it,we have encoded the object type data by LabelEncoder so now data is more simplified for further analysis lets will check the total number of column present in data set

Data contains 71 columns and 269 rows

4)checking null values:

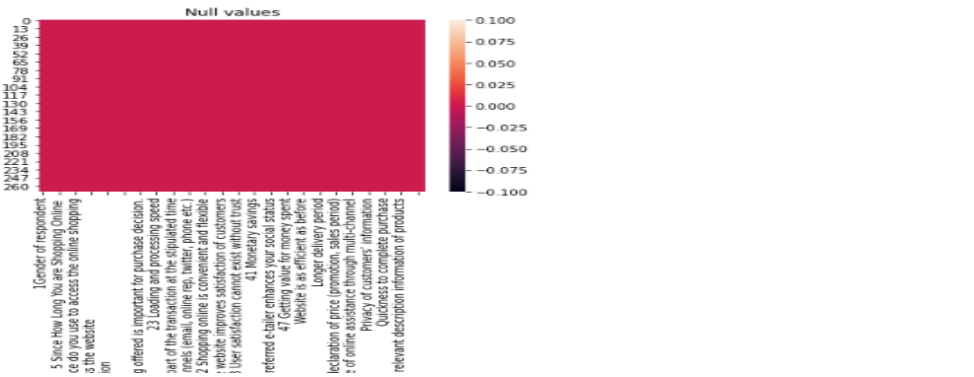
In this case we have to find out the null values present in our data set if yes it is required to remove it.in our data set it does not have any null values it is also shown by heatmap

```
In [33]: data.isnull().sum().sum()
Out[33]: 0
```

no NULL values present in our data set for better understanding we can also represent it with help of heatmap

```
In [34]: import seaborn as sns

In [35]: sns.heatmap(data.isnull())
         plt.title("Null values")
         plt.show()
```



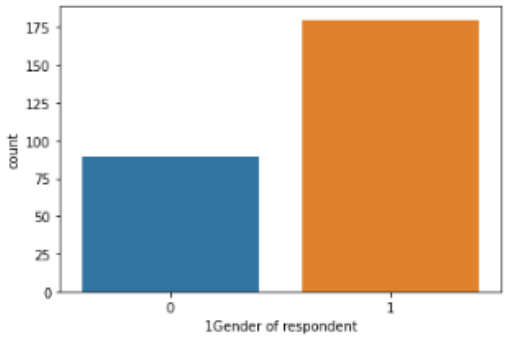
5) visualization:

visualization

```
In [36]: data_visualization=data[['1Gender of respondent']].copy()
```

```
In [37]: import seaborn as sns
         ax=sns.countplot(x='1Gender of respondent',data=data_visualization)
         print(data_visualization['1Gender of respondent'].value_counts())
```

```
1    180
0     89
Name: 1Gender of respondent, dtype: int64
```

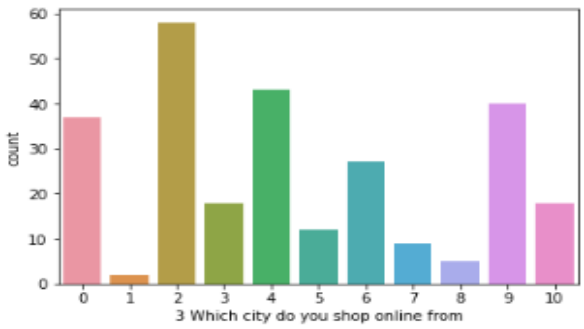


the gender of respondant is 180 female and 89 are male out of 269 so %of female respondent is more as compare to male

As the count plot shows distribution of data in respective column in male female category.

```
In [39]: import seaborn as sns
         ax=sns.countplot(x='3 Which city do you shop online from',data=data_visualization1)
         print(data_visualization1['3 Which city do you shop online from'].value_counts())
```

```
2     58
4     43
9     40
0     37
6     27
3     18
10    18
5     12
7      9
8      5
1      2
Name: 3 Which city do you shop online from, dtype: int64
```



city from which maximum % of online shopping done is city no 2

It shows that in which city or from which city online shopping is done

```
In [41]: import seaborn as sns
ax=sns.countplot(x='6 How many times you have made an online purchase in the past 1 year',data=data_visualization2)
print(data_visualization2['6 How many times you have made an online purchase in the past 1 year'].value_counts())
```

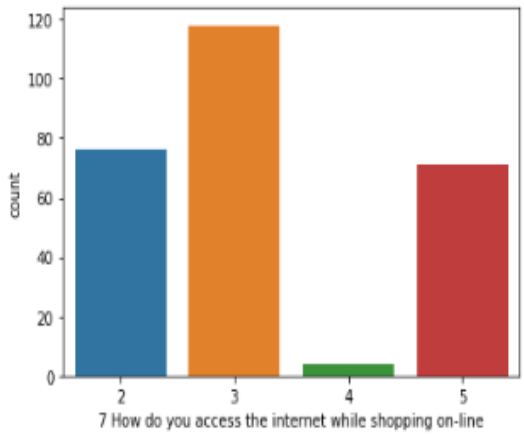
```
1    114
4     63
5     53
2     29
3     10
Name: 6 How many times you have made an online purchase in the past 1 year, dtype: int64
```

so from category 1 we got maximum purchase

It shows that how many times customer has made purchase in past one year

```
In [43]: import seaborn as sns
ax=sns.countplot(x='7 How do you access the internet while shopping on-line',data=data_visualization3)
print(data_visualization3['7 How do you access the internet while shopping on-line'].value_counts())
```

```
3    118
2     76
5     71
4      4
Name: 7 How do you access the internet while shopping on-line, dtype: int64
```



while shopping online maximum access carried out through the category 3 i.e 118

It shows that how the customer access is there while doing online shopping

In [45]:
```
import seaborn as sns
ax=sns.countplot(x='8 Which device do you use to access the online shopping',data=data_visualization4)
print(data_visualization4['8 Which device do you use to access the online shopping'].value_counts())
```

```
1    141
2     86
3     30
4     12
Name: 8 Which device do you use to access the online shopping, dtype: int64
```

device used to access the online shopping is from category 1 it is about 141

It clearly shows that which device the customer were using while doing the inline shopping maximum coming under category 1

6) Data Description:

In this case data is described in detail which helping us for detail analysis

now lets will describe our data set for detail analysis

In [46]:
```
data.describe()
```

Out[46]:

	1Gender of respondent	2 How old are you	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	10 What is the operating system (OS) of your device\t\t	11 What browser do you run on your device to access the website	...	Speedy order delivery	Availability of several payment options
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	2.959108	220465.747212	3.524164	2.672862	3.260223	1.676580	4.282528	1.776952	1.275093	...	1.301115	3.680297
std	0.471398	1.066012	140524.341051	1.436586	1.651788	1.135887	0.843904	0.923426	0.797892	0.645429	...	1.497024	3.043531
min	0.000000	1.000000	110008.000000	1.000000	1.000000	2.000000	1.000000	2.000000	1.000000	1.000000	...	0.000000	0.000000
25%	0.000000	2.000000	122018.000000	3.000000	1.000000	2.000000	1.000000	4.000000	1.000000	1.000000	...	0.000000	1.000000
50%	1.000000	3.000000	201303.000000	4.000000	2.000000	3.000000	1.000000	4.000000	2.000000	1.000000	...	1.000000	3.000000
75%	1.000000	4.000000	201310.000000	5.000000	4.000000	5.000000	2.000000	5.000000	2.000000	1.000000	...	2.000000	5.000000
max	1.000000	5.000000	560037.000000	5.000000	5.000000	5.000000	4.000000	5.000000	3.000000	4.000000	...	5.000000	10.000000

8 rows × 71 columns

it shows total count of every column that every column contains 269 values, also it shows mean and std of every column, we observe that value of std is low as compare to the mean value in every column it means that our dataset is good for analysis also every column show its min,25%,50%,75%,max data values in it.

7) Data Correlation:

```
In [47]: data.corr()
```

Out[47]:

	1Gender of respondent	2 How old are you	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	10 What is the operating system (OS) of your device\t\t	11 What browser do you run on your device to access the website	...	Speedy order delivery	Availability of several payment options	Quickness: t complet purchas
1Gender of respondent	1.000000	-0.034449	0.260696	-0.046005	0.076121	-0.200974	-0.241847	-0.058760	-0.038207	-0.116712	...	-0.085661	-0.050594	-0.06010
2 How old are you	-0.034449	1.000000	-0.057393	0.014049	0.005089	0.024228	0.292176	0.227841	-0.133597	-0.151708	...	-0.169956	-0.082249	-0.15304
4 What is the Pin Code of where you shop online from	0.260696	-0.057393	1.000000	0.092980	-0.021229	0.134879	-0.103424	-0.189814	0.077905	0.063588	...	-0.172836	-0.046428	0.02467
5 Since How Long You are Shopping Online	-0.046005	0.014049	0.092980	1.000000	0.291104	0.080738	0.309634	0.118596	-0.141768	-0.236578	...	-0.051109	0.116130	0.04450
6 How many times you have made an online purchase in the past 1 year	0.076121	0.005089	-0.021229	0.291104	1.000000	0.135034	0.057656	-0.127544	0.040689	-0.153269	...	-0.154673	0.218856	0.14716
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Complete, relevant description information of products	-0.124719	0.038227	0.010051	0.058788	0.188406	0.056179	0.177311	-0.035103	0.153374	-0.214545	...	-0.007663	0.801683	0.65781

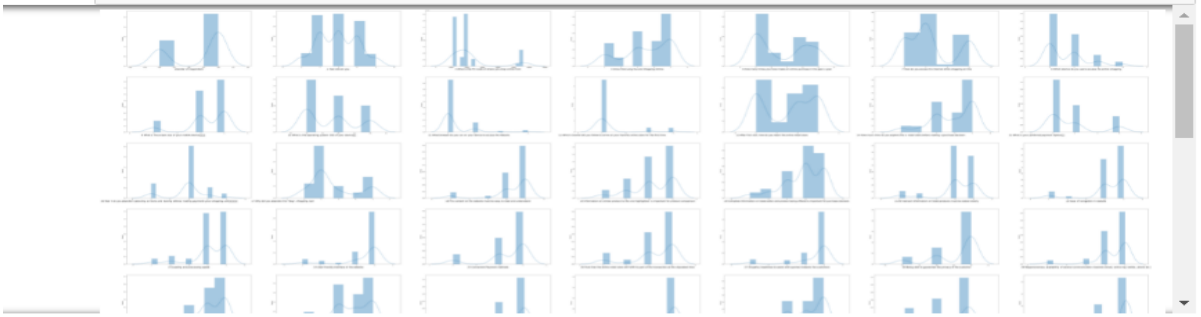
Data is correlated with other column data and also with its own it also gives the positive negative correlation of data with respective one another

8) Data Distribution(subplot):

It shows how data is distributed in each column

now will plot subplot to know how data is distributed in every column

```
In [49]: #now lets see how data is distributed in each column
plt.figure(figsize=(150,130),facecolor='white')
plotnumber=1
for column in data:
    if plotnumber<=71:
        ax=plt.subplot(11,7,plotnumber)
        sns.distplot(data[column])
        plt.xlabel(column,fontsize=20)
    plotnumber+=1
plt.show()
```

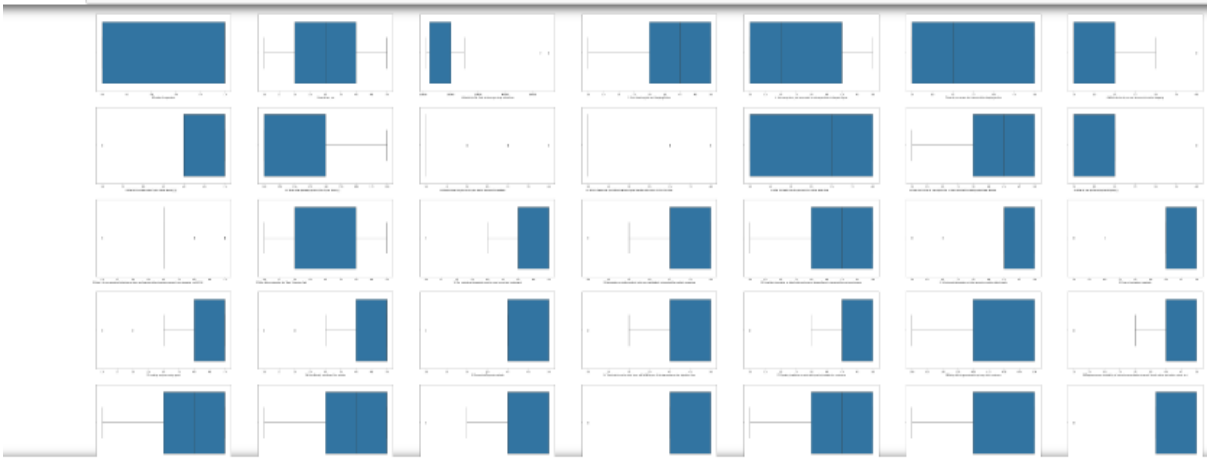


subplot shows that some column data contains skewness which is required to remove and some are uniformely distributed,we also plot histogram for better understanding of how data is distributed in every column

It shows some column have uniformely distributed data some have skewness

9)Detection of outliers:

```
In [52]: #visualize the outliers using boxplot
         plt.figure(figsize=(100,80))
         graph=1
         for column in data:
             if graph<71:
                 ax=plt.subplot(11,7,graph)
                 sns.boxplot(data[column],orient='V')
                 plt.xlabel(column,fontsize=10)
                 graph+=1
         plt.show()
```



some column contains outliers which is required to remove,will remove it in further steps

Some column has outliers present in it is necessary to remove that outliers that can be removed with zscore test and by selecting appropriate threshold values

9) Checking Skewness:in this case we have checked whether any skewness present in data or not some column has skewness in it we have removed it

10) Removing outliers:

now we will remove the outliers

In [54]:
```
#calculate the zscore
z = np.abs(zscore(data))
print(z)
```

```
[[1.42213639 0.03843148 0.78749796 ... 1.4341911  1.47052245 2.29324079]
 [0.70316744 0.9013929  0.78734824 ... 0.72107205 0.82131149 0.19585849]
 [0.70316744 0.9013929  0.13658457 ... 1.07763158 1.92888924 1.46354103]
 ...
 [0.70316744 0.97825586 2.42077023 ... 1.06172557 1.01215566 1.44040814]
 [0.70316744 1.84121727 0.33677293 ... 0.34860652 0.82131149 0.19585849]
 [0.70316744 0.97825586 0.13871628 ... 1.06172557 1.73804507 1.44040814]]
```

In [55]:
```
threshold=3
print(np.where(z<3))
print(data.shape)
```

```
(array([  0,   0,   0, ..., 268, 268, 268], dtype=int64), array([ 0,  1,  2, ..., 68, 69, 70], dtype=int64))
(269, 71)
```

In [56]:
```
#Assign the value to df_new which are less the threshold value and removing the outliers
data_new=data[(z<3).all(axis = 1)]
```

In [57]:
```
print(data.shape)
print(data_new.shape)
data = data_new
print('Shape after removing outlires',data.shape)
```

```
(269, 71)
(215, 71)
Shape after removing outlires (215, 71)
```

so after removing the outliers we have 188 rows remain and 71 columns , now again will plot subplot to check how data is distributed after removing the outliers

So outliers are removed from data after removing outliers we have 215 rows and 71 column remains

CONCLUSION

We observed that data was filled with some outliers it is removed, also there some encoding is done now data is uniformly distributed in column. it is also observed from subplot, so now by selecting suitable algorithm we can use this data for building machine learning algorithm for predicting the future data