A
Project on

CUSTOMER RETENTION CASE STUDY

By
Snehal Kuamwat

Content

- Introduction
- Data preparation & cleaning
- Data analysis
- Regressor Model Building
- Limitations of this work and Scope for Future Work

Introduction

- **Business Problem Framing**

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

- **Conceptual Background of the Domain Problem**

  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.

- **Review of Literature**

  A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention.

- **Motivation for the Problem Undertaken**

  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.

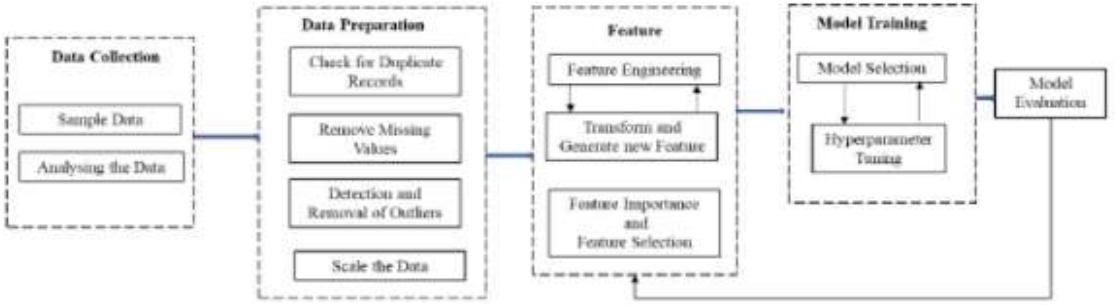
- **Data Sources and their formats**

  The dataset was in the form of excel file, so I used the read excel file function from the pandas module. The picture of the dataset I just have given below, you can observe that it consists of 269 rows (records) and 71 columns (features).

	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	...	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 years	Greater Noida	201308	3-4 years	41 times and above	Mobile Internet	Smartphone	5.5 inches	Android	...	Myntra.com	Myntra.com	Myntra.com
3	Male	21-30 years	Karnal	132001	3-4 years	Less than 10 times	Mobile Internet	Smartphone	5.5 inches	IOS/Mac	...	Snapdeal.com	Myntra.com, Snapdeal.com	Myntra.com
4	Female	21-30 years	Bangalore	530068	2-3 years	11-20 times	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	...	Flipkart.com, Paytm.com	Paytm.com	Paytm.com
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
264	Female	21-30 years	Solan	173212	1-2 years	Less than 10 times	Mobile Internet	Smartphone	5.5 inches	Android	...	Amazon.in	Amazon.in	Amazon.in
265	Female	31-40 years	Ghaziabad	201008	1-2 years	31-40 times	Mobile Internet	Smartphone	Others	Android	...	Flipkart.com	Flipkart.com	Flipkart.com
266	Female	41-50	Bangalore	560010	2-3 years	Less than	Mobile	Laptop	Others	Window/windows		Amazon.in	Snapdeal.com	Amazon.in

1. Data preparation and cleaning

Fig. 1 captures the research framework for the housing price prediction problem. It includes five major blocks, namely data collection, data preparation, feature processing, mode training, and model evaluation. These blocks of the diagram are explained in detail in the next subsections.



2. Data Analysis

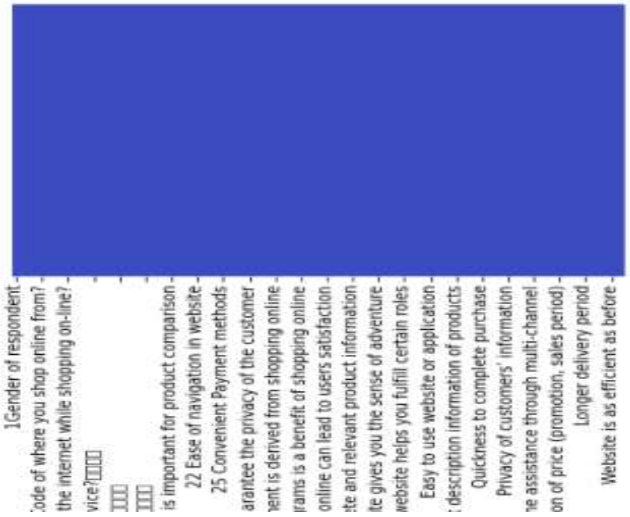
In this we need to extract data from excel file. After understanding the datasets we need to clean the data if some features has missing values. Let's just remove the features with 30% or less NaN values.

But, this dataset has not a NaN values in it.

Checking the null values

```
In [49]: import seaborn as sns
         sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='coolwarm')
```

```
Out[49]: <AxesSubplot:>
```



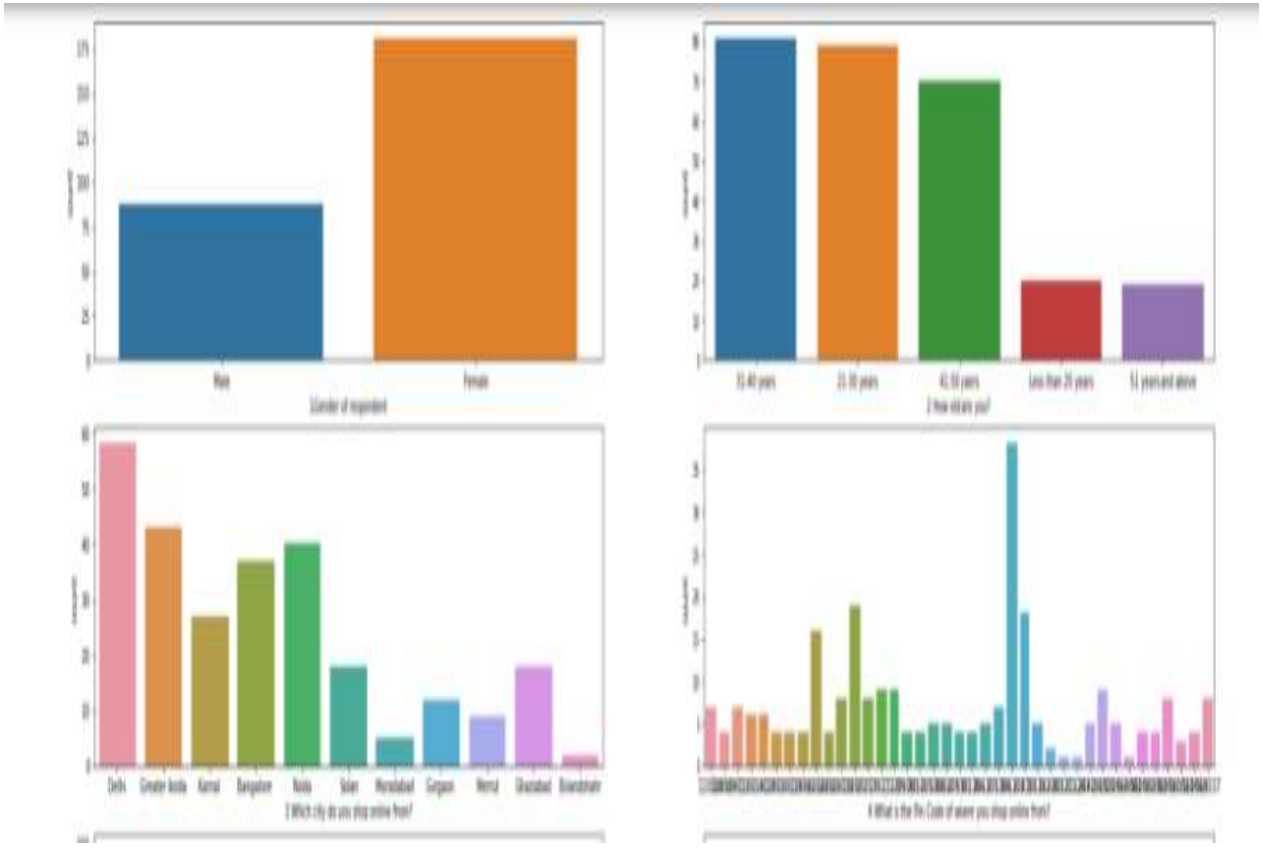
Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Code

```
import matplotlib.pyplot as plt
rows = 36
columns = 2
fig, axes = plt.subplots(rows,columns, figsize=(30,150))
x, y = 0, 0
for i, column in enumerate(df.columns):
    sns.countplot(x=df[column], ax=axes[x, y])
        if y < columns-1:
        y += 1
    elif y == columns-1:
        x += 1
        y = 0
    else:
        y += 1
```

Output



Encoding of the dataset

Code

```
from sklearn.preprocessing import LabelEncoder
labellencoder=LabelEncoder()
for column in df.columns:
    df[column]=labellencoder.fit_transform(df[column])
df
```

	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)	Longer page loading time (promotion, sales period)	U
0	1	1	2	1	3	2	0	0	3	2	...	0	0	3	5	
1	0	0	2	5	3	3	3	2	0	1	...	1	6	7	10	
2	0	0	4	23	2	3	1	2	2	0	...	7	6	4	7	
3	1	0	6	11	2	5	1	2	2	1	...	9	7	4	8	
4	0	0	0	31	1	0	3	2	0	1	...	5	8	5	8	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
264	0	0	10	13	0	5	1	2	2	0	...	0	0	0	0	
265	0	1	3	17	0	2	1	2	3	0	...	4	4	3	5	
266	0	2	0	35	1	5	2	1	3	2	...	0	9	0	10	
267	0	4	10	14	1	5	3	2	2	0	...	0	2	0	4	
268	0	2	3	18	1	2	1	2	2	0	...	0	0	0	0	

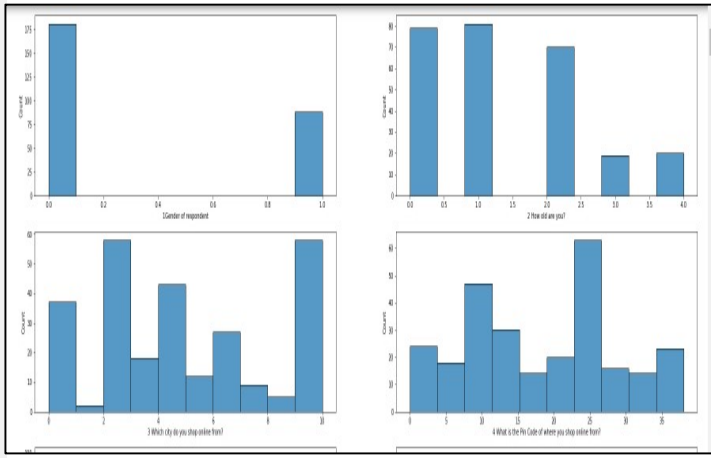
269 rows × 71 columns

Data Visualization

Code

```
import matplotlib.pyplot as plt
rows = 36
columns = 2
fig, axes = plt.subplots(rows,columns, figsize=(30,180))
x, y = 0, 0
for i, column in enumerate(df.columns):
    sns.histplot(x=df[column], ax=axes[x, y])
        if y < columns-1:
        y += 1
    elif y == columns-1:
        x += 1
        y = 0
    else:
        y += 1
```

Output



All the columns are nonzero values.

After that we checking the Outliers and then checking correlation of features.

using df.corr( ) and the output is

	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)
1Gender of respondent	1.000000	0.046169	0.080912	-0.289628	-0.057096	0.077876	-0.309029	0.061673	0.028794	-0.019243	...	-0.101925	-0.228744
2 How old are you?	0.046169	1.000000	0.113712	-0.133946	-0.087847	0.309575	0.255594	0.022383	-0.006101	-0.048087	...	-0.281877	-0.095850
3 Which city do you shop online from?	0.080912	0.113712	1.000000	-0.064136	-0.138329	0.173871	-0.010436	0.020650	0.199296	-0.051642	...	-0.065450	-0.115453
4 What is the Pin Code of where you shop online from?	-0.289628	-0.133946	-0.064136	1.000000	-0.074280	-0.304554	-0.035490	-0.021647	-0.116924	-0.051971	...	0.003349	0.128535
5 Since How Long You are Shopping Online ?	-0.057096	-0.087847	-0.138329	-0.074280	1.000000	0.013315	0.226883	-0.125240	0.139924	0.274201	...	0.154763	-0.063386
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Longer delivery period	0.060838	-0.156173	-0.123369	-0.076998	0.218641	-0.130651	0.101297	-0.104665	0.048533	0.397953	...	0.268484	0.365030
Change in website/Application design	-0.164818	-0.134558	0.000427	0.001954	0.220347	0.007841	0.147770	-0.052146	0.418180	0.059894	...	0.370649	0.152655
Frequent disruption when moving from one page to another	-0.256638	-0.018825	0.019167	0.113557	0.025919	-0.127148	0.349813	0.084876	-0.065658	0.111380	...	0.274454	0.603750
Website is as efficient as before	0.055663	-0.008582	0.007117	-0.008289	-0.024316	-0.124076	0.266932	0.333868	-0.100462	-0.128611	...	0.122555	-0.010967
Which of the Indian online retailer would you recommend to a friend?	-0.003372	-0.135263	-0.142123	-0.097320	0.136106	-0.152028	0.041129	0.099425	0.074453	-0.159579	...	0.261774	-0.140519

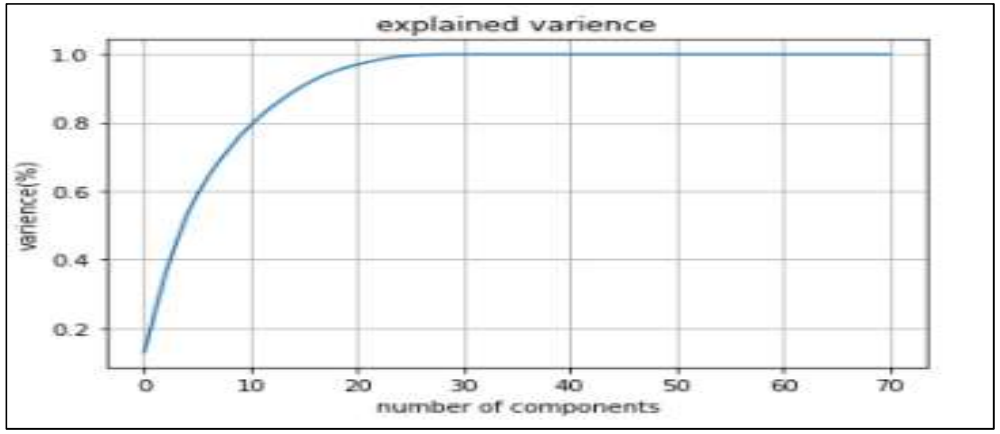
Then we checked skewness of the data. And the threshold skewness is +/-0.5, hence the skewness values are above and below the threshold value hence we remove the skewness. After removing the skewness, we scaled the data.

After the scaling data, the data contains 71 column and having muliti collinearity to remove it we need to use Principal Component Analysis (PCA).

Code

```
from sklearn.decomposition import PCA
pca=PCA()
pc=pca.fit_transform(x_t)
plt.figure()
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel("number of components")
plt.ylabel("varience(%)")
plt.title("explained varience")
plt.show()
```

Output



So now we selected 30 columns hence principal_x is having 30 columns and 249 rows.

3. Regressor Model Building

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

4. Limitations of this work and Scope for Future Work

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.