CUSTOMER RETENTION ANALYSIS

Data analysis is a process of finding, collecting, cleaning, examining, and modeling data to derive useful information and insights and understand the derived information for data-driven decision-making.

Why Data Analysis is Needed?

Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analysis have been automated into mechanical processes and algorithms that work over raw data for human consumption.

Problem Statement

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
- 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.

Steps Followed in Data analysis

- Define the business objective.
- Source and collection of data.
- Processing and cleaning the data.
- Perform exploratory data analysis (EDA).
- Select, build, and test models.
- Monitor and validate against stated objectives.

Objective

The main objective of Customer retention is to find Which are the websites that most people recommend to others depending upon the customers satisfaction and trust on the particular site.

Source and collection of data.

- Data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.
- Some common sources or methods of collecting primary data are:
  - Interviews
  - surveys
  - questionnaires
  - experiments
  - observations.

Here the data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

Processing of the data

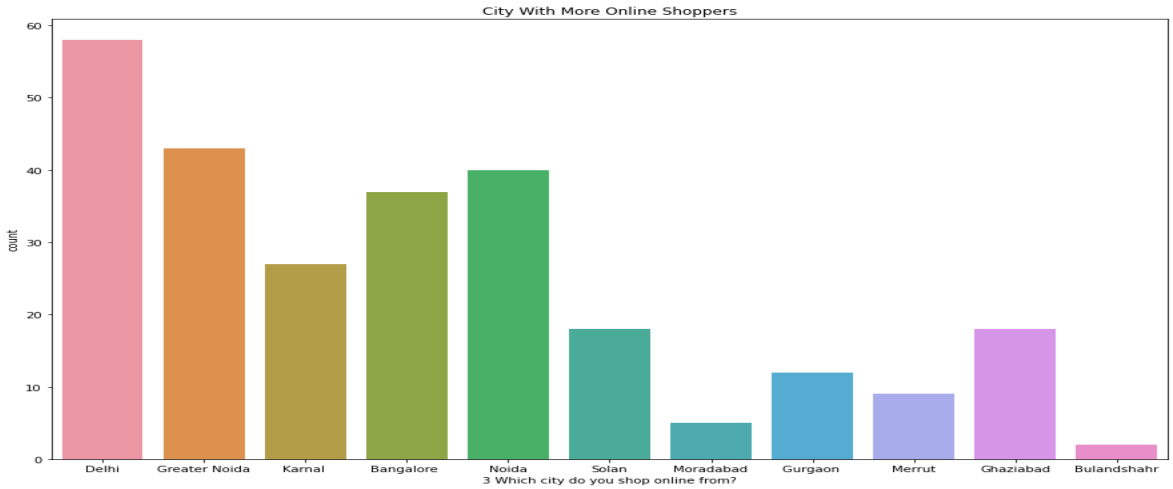
First thing is problem statement analysis, statistics & Eda. Before jumping to machine learning its always encouraged to get a good grasp over the problem cause the better understanding it, the better we'll form the end goal hence better the result.

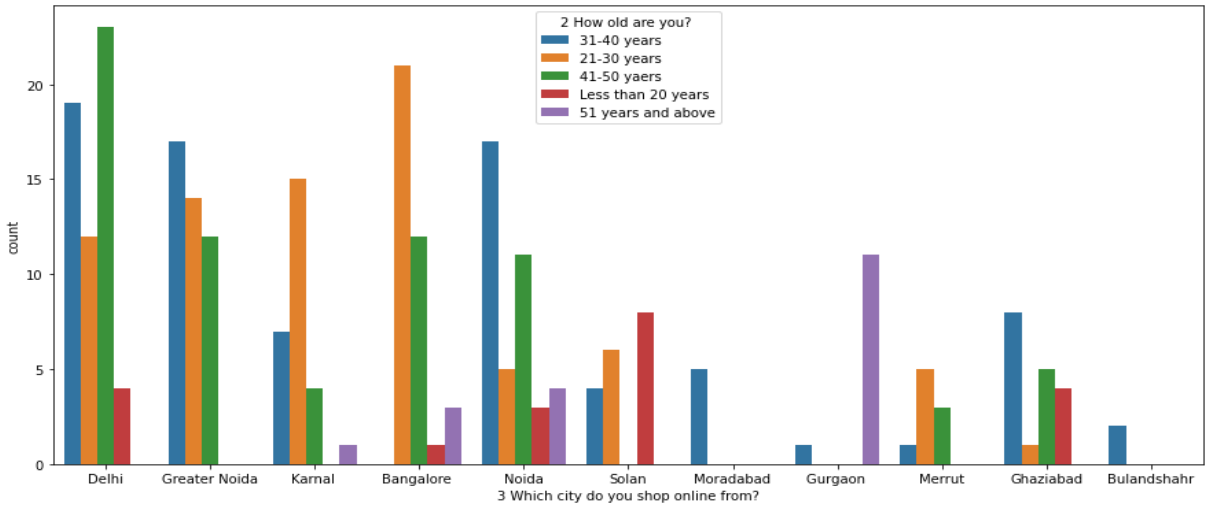
I started solving the problem by checking whether the dataset have any null values or not. But, there are no null values in the dataset.

Now Comes the EDA into picture, Doing EDA is always crucial for the end product as it gives huge insights into the data that cannot be acheived while looking at the tabular data.

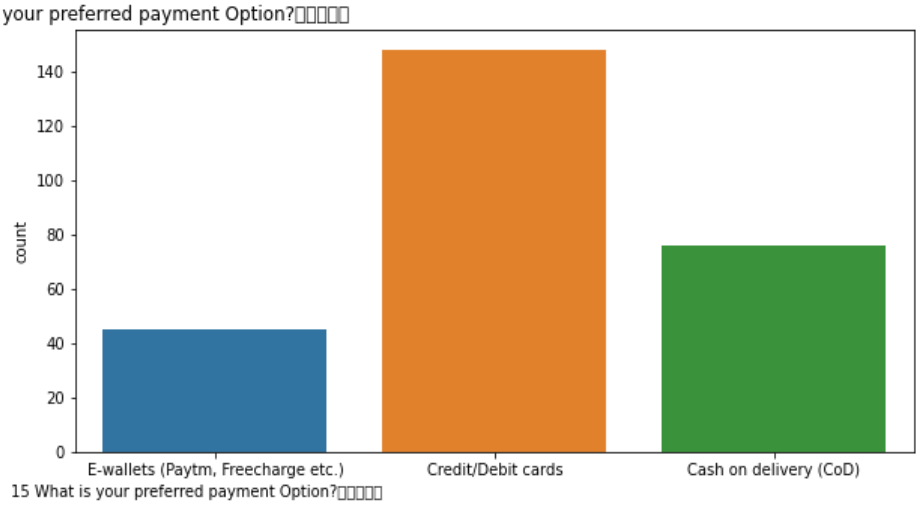
I've plotted several countplots, heatmap & a boxplot to count, measure, relate & most importantly visualize the data.

Some visualizations:

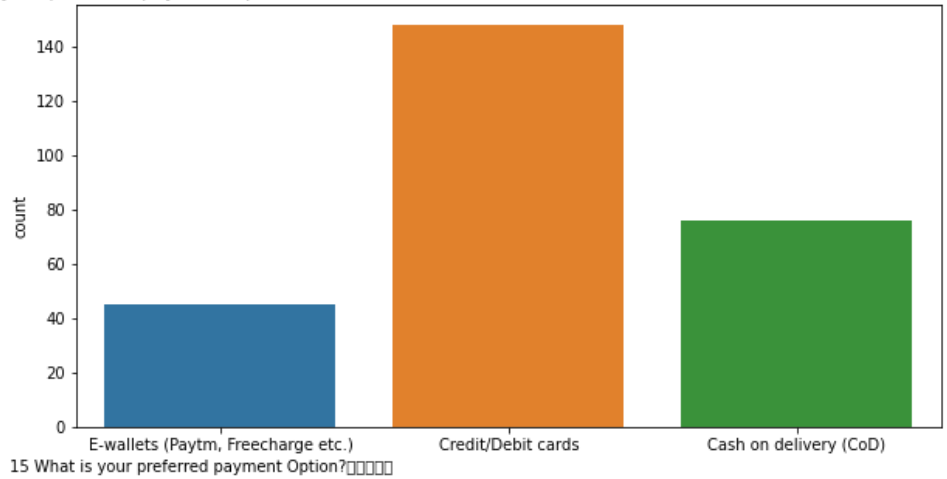


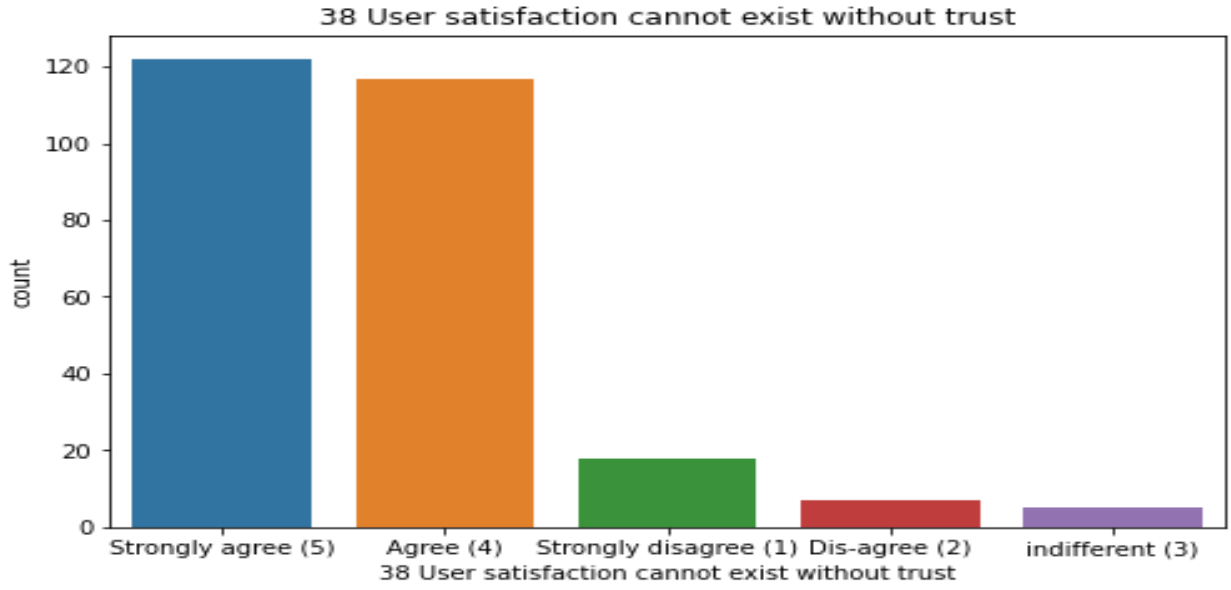


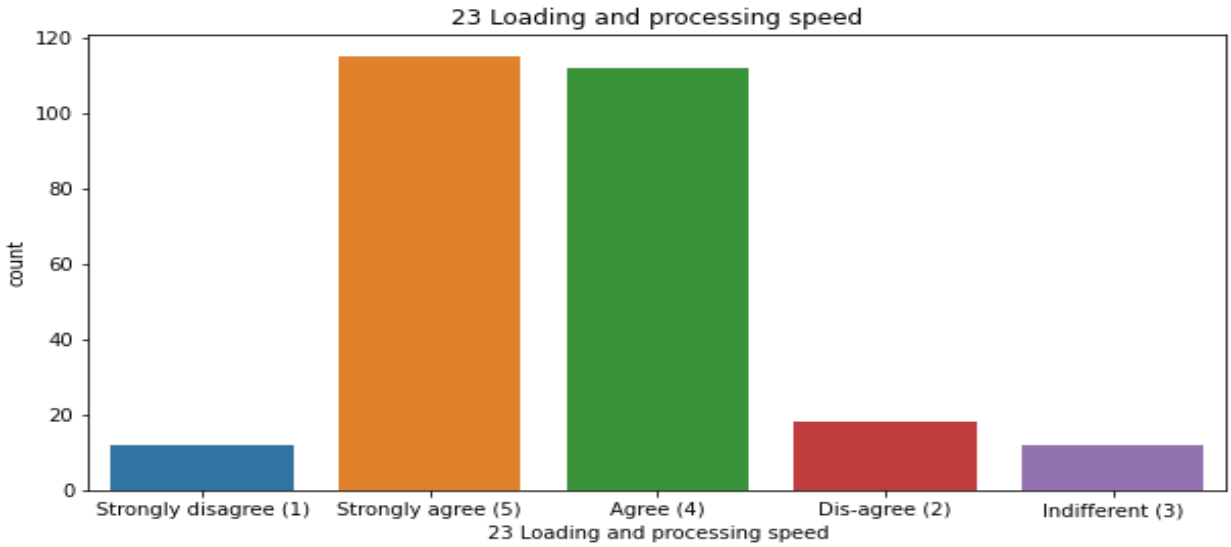
15 What is your preferred payment Option?□□□□□

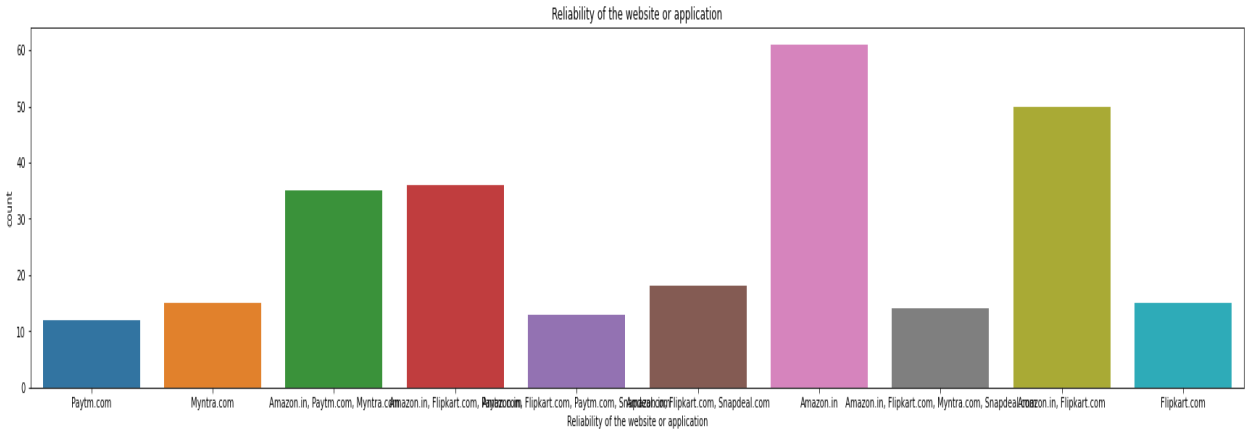


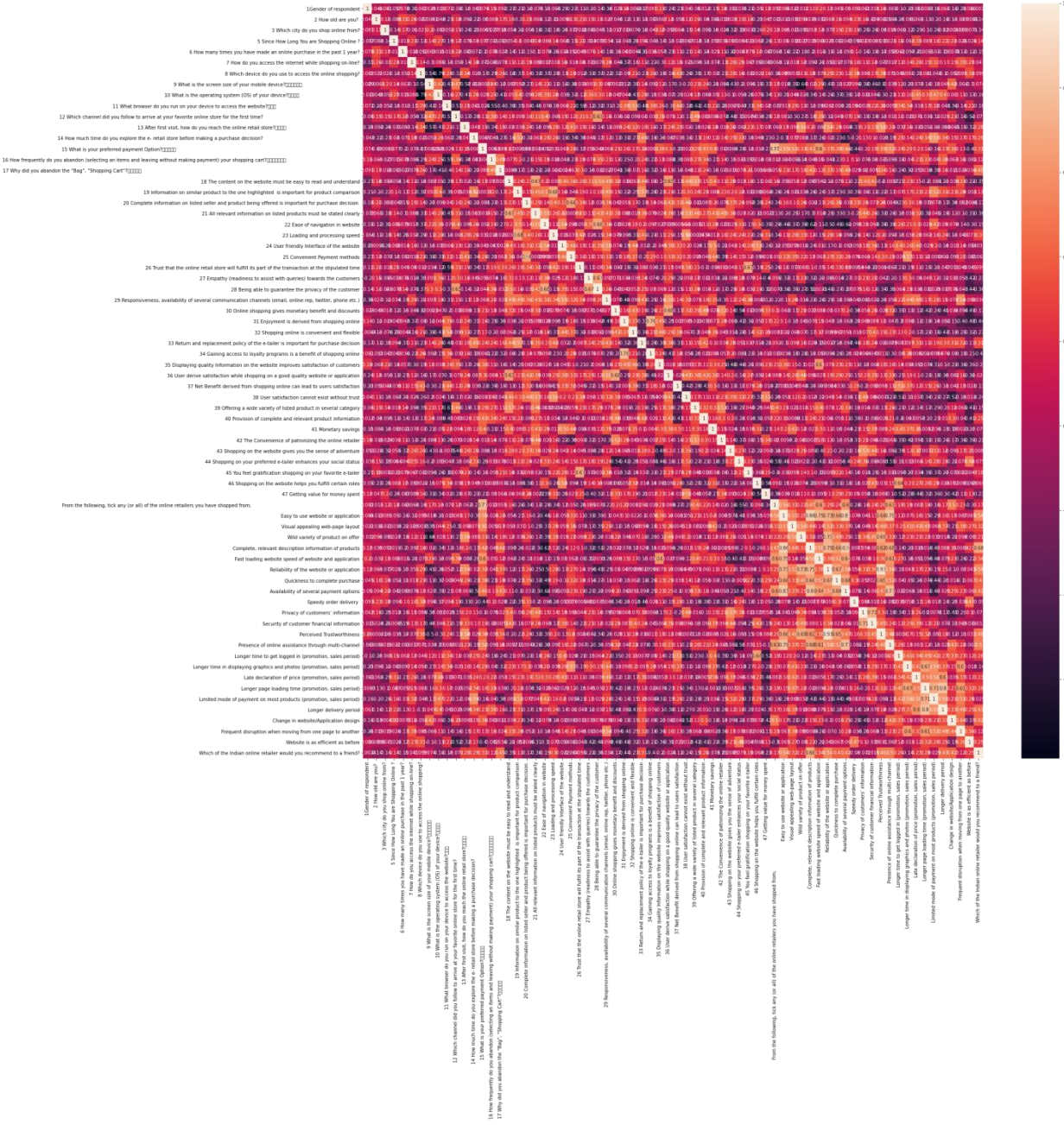
15 What is your preferred payment Option?□□□□□




38 User satisfaction cannot exist without trust
count
120
100
80
60
40
20
0
Strongly agree (5)
Agree (4)
Strongly disagree (1)
Dis-agree (2)
indifferent (3)
38 User satisfaction cannot exist without trust


23 Loading and processing speed
count
120
100
80
60
40
20
0
Strongly disagree (1)
Strongly agree (5)
Agree (4)
Dis-agree (2)
Indifferent (3)
23 Loading and processing speed


Reliability of the website or application
count
60
50
40
30
20
10
0
Paytm.com
Myntra.com
Amazon.in, Paytm.com, Myntra.com
Amazon.in, Flipkart.com, Paytm.com
Flipkart.com, Paytm.com, Snapdeal.com
Flipkart.com, Snapdeal.com
Amazon.in
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com
Amazon.in, Flipkart.com
Flipkart.com
Reliability of the website or application



Some key observations:

- City with more online shoppers are seen from Delhi and least shoppers are from Bulandshahr.

- This shows the different age group of shoppers from various cities. People from age group of 21-30 prefer more online shopping.
- For the past 1 year, most people have purchased less than 10 times.
- Most shoppers use MOBILE INTERNET on SMART PHONES to make online purchase.
- Widely used browser for online shopping is GOOGLE CHROME on WINDOWS
- From the above observations, SEARCH ENGINES are the most favourite channel for online shoppers, even after shopping for the first time.
- Before making an order, people search for more than 15 mins to fix it and the preferred method of payment is through CARD.
- With effective ALTERNATIVE OFFERS people tend to ABANDOM the cart SOMETIMES.
- People strongly recommend readability and understandability of the products and information on similar products and sellers for making a decision to purchase online.

Frequent Online Shoppers strongly recommend:

- Ease of Navigation
- High processing speed
- user friendly
- convenient Payment
- Empathy
- enjoyment from online shopping

Most people strongly recommend:

- Convenient Shopping
- with good returns policy
- User satisfaction with trust
- Monetary Savings.

From the above insights, certain websites are being recommended by the customers based on different criteria. Mostly **amazon.in** and other combined websites.

Data Description

First, the categorical data is encoded to make it to numerical type.

Description function provides the values like

- Mean
- Median
- Mode
- Maximum values
- Minimum values

In certain columns, the mean values are deviated much from the median showing the presence of skewness.

In some columns the values of 75th percentile and maximum values are more indicating the presence of outliers.

The Correlation within each column is found and also a with the target variable is found.

Then the columns with multicollinearity is found using *Variance Inflation Factor*. Those columns that are closely related to eachother are removed based on its least contribution to the target variable. Many columns have high collinearity within them. So, certain columns are deleted based on its lowest contribution to the target variable.

Skewness and Outliers Removal

After removing the unnecessary columns the dataset is checked for skewness by using **Power transform** method, and is removed to make it suitable for training the

model. After removal of skewness, the dataset is scaled to avoid large gap between the values. The outliers are removed by using interquartile method. Now the dataset is ready to train and test.

Finally, model building & prototyping comes into play, I prefer to use a model to get good predictive scores for this problem. As the data is of higher dimension i.e. significant no of features, I think it's better to use high variance & low bias models like Decision Tree. So In order to build a better predictive model, I think its a good practice is to adopt the ensemble methods. The model provided the best result.

Then, I went for Parameter tuning with Grid Search CV.With the best parameter and estimator the results are verified. I also tested the model with Lasso Regularization to confirm the model is not overfitting to the dataset.

The model is saved with pickle library.

The model is trained and tested to find the accuracy. The model with Decision Tree Classifier gave 100% accuracy which implies the model is trained well. The dataset works well with the selected model. The most recommended website is amazon.com.