PROJECT REPORT

CUSTOMER RETENTION

Submitted by:

UJJWAL PRATIK

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Faculty and SME for the guidance. They have covered the topics like Machine Language, Python & SQL. I had also taken help from YouTube & online videos.

Customer Retention Analysis

Customer retention analysis is the application of statistics to understand how long customers are retained before churning out and to identify trends in customer retention. This type of analysis discerns how long customers usually stick around, whether seasonality affects customer retention, and discovers behaviours and factors that differentiate retained customers from churned customers.

Importance of Customer Retention Analysis

Customer retention analysis is important for any company because it helps you understand which personas have higher retention rates and discern which features impact retention. This provides actionable insights that can help you make more effective product and marketing decisions. It can be difficult for a product or sales team to know how well a product is performing with the target audience. They may think that features and messaging is on brand and clear because acquisition numbers are growing. However, just because new customers are purchasing a product does not necessarily mean customers like the product or service enough to stick around. That is where customer retention analytics comes in. Every company needs data to make effective business and marketing decisions. Machine learning makes this easier than it has ever been before, which is great news for companies that wish to leverage this data.

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 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.

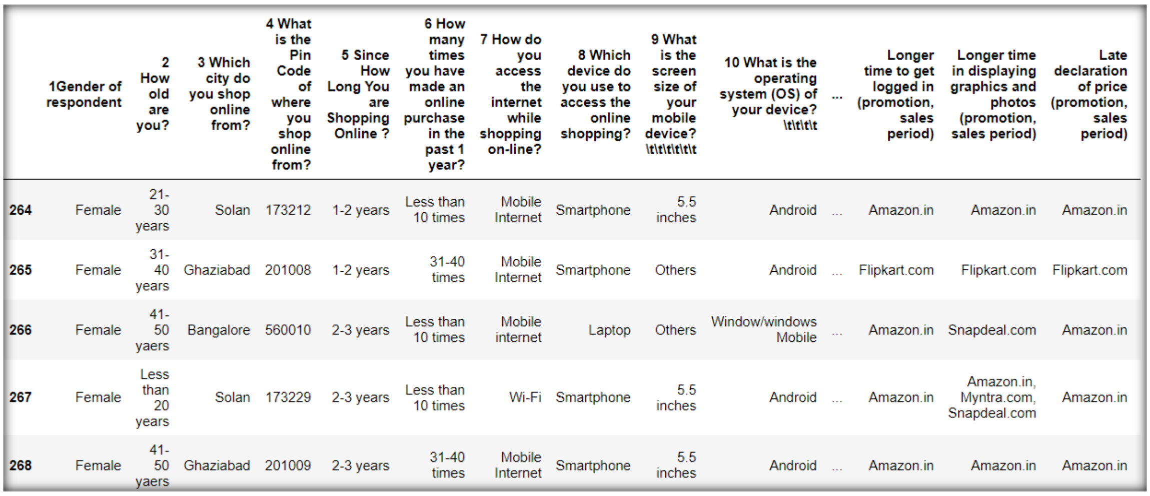
Objective

- Our main objective is to analyse the data using machine learning techniques.
- We use physical characteristic to predict age using learning algorithms.

Data Description

- The dataset contains 71 columns and 269 rows
- 70 object data type columns and 1 integer data type column

Data Samples in Panda



	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)
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 yaers	Bangalore	560010	2-3 years	Less than 10 times	Mobile internet	Laptop	Others	Window/windows Mobile	...	Amazon.in	Snapdeal.com	Amazon.in
267	Female	Less than 20 years	Solan	173229	2-3 years	Less than 10 times	Wi-Fi	Smartphone	5.5 inches	Android	...	Amazon.in	Amazon.in, Myntra.com, Snapdeal.com	Amazon.in
268	Female	41-50 yaers	Ghaziabad	201009	2-3 years	31-40 times	Mobile Internet	Smartphone	5.5 inches	Android	...	Amazon.in	Amazon.in	Amazon.in

Data Analysis

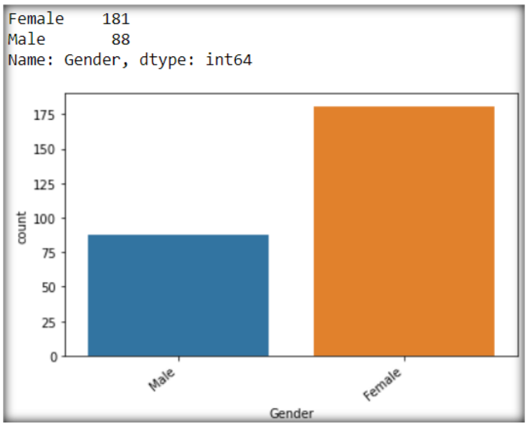
In the first stage of the project, it's very important to analyse and perform data analysis. We run statistical analysis of all available attributes, analyse existing data structure, as well as customer care department actions and all related business aspects. Following analysis is done as part of this project.

- Analysis of available data types
- Visual data analysis
- Correlation analysis
- Missing values analysis
- Analysis and definition of the "target" variable

While analysing the data, the features which influence people to do online shopping are many. For example, Females are more keen to do online shopping than male. It is preferred in the metro than non-metro or smaller town. Age is also seen as key influencing factor as elderly people have shown lesser interest than Young and mid aged people. This is also influenced by Smart mobile device which helps such age group of people for online shopping.

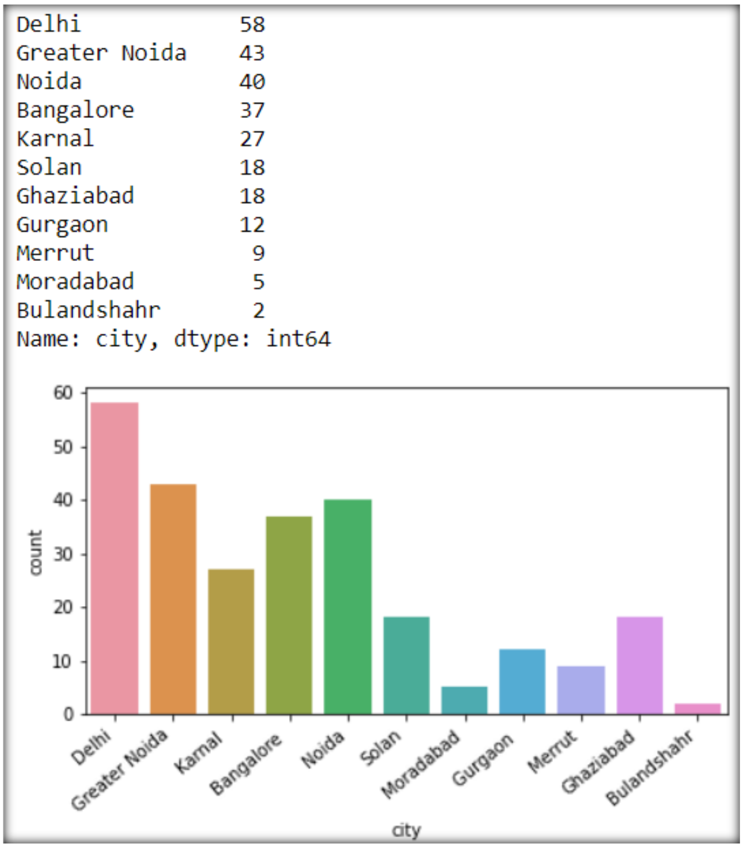
With advance graphics feature of Python using Seaborn, all these analysis can be translated into visualizations.

For example, visualization of interests of male vs female in online shopping is:



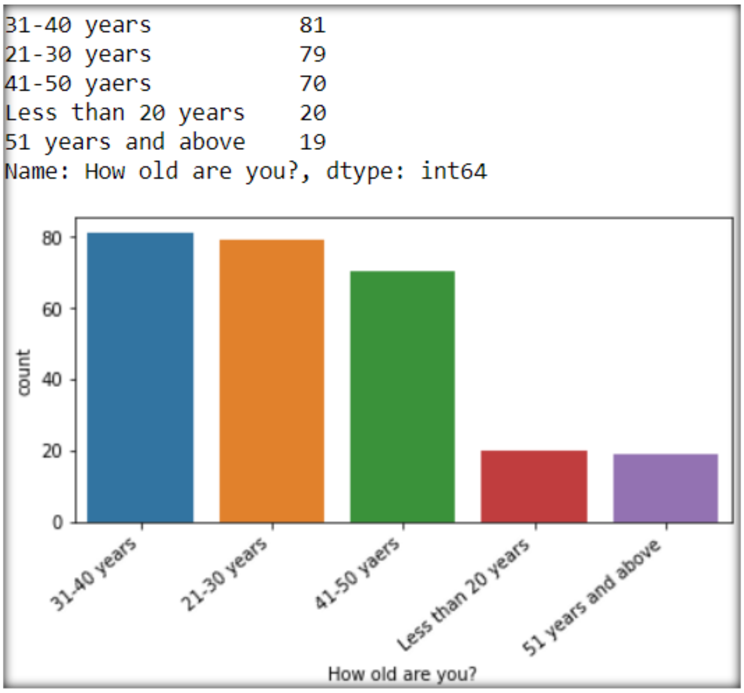
```
Female    181
Male       88
Name: Gender, dtype: int64
```

Metro vs Non-Metro users can be seen as:



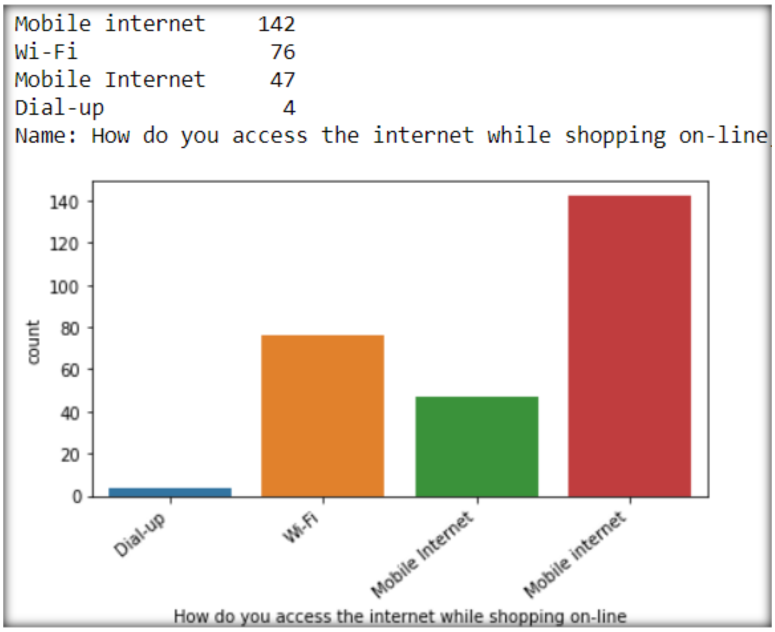
```
Delhi            58
Greater Noida    43
Noida            40
Bangalore        37
Karnal           27
Solan            18
Ghaziabad        18
Gurgaon          12
Merrut            9
Moradabad         5
Bulandshahr       2
Name: city, dtype: int64
```

Young and Elderly people’s interest for online shopping is:



```
31-40 years           81
21-30 years           79
41-50 yaers           70
Less than 20 years    20
51 years and above    19
Name: How old are you?, dtype: int64
```

Infographics of shopping with Smart mobile devices is:

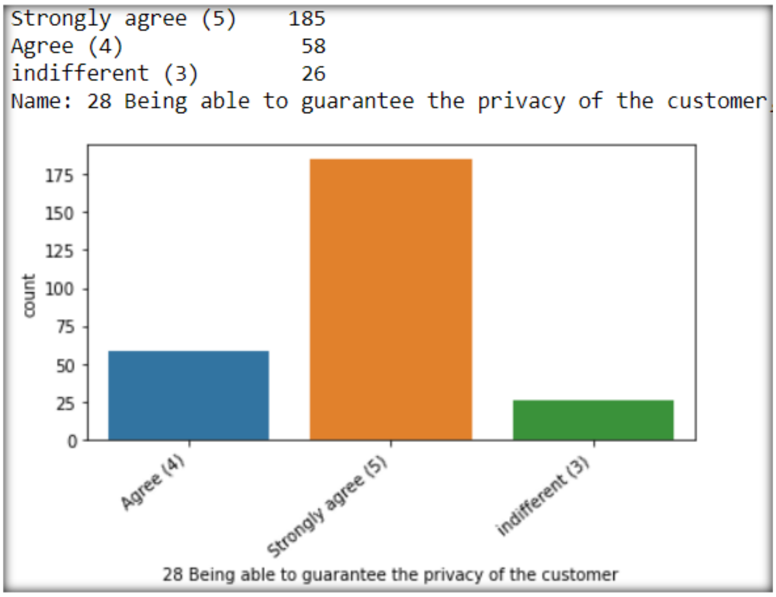


```
Mobile internet    142
Wi-Fi               76
Mobile Internet     47
Dial-up              4
Name: How do you access the internet while shopping on-line
```

Further analysis of the data shown that people are more interested to shop from the website/vender who usually provides product ratings and comparisons. Ease of navigation, Site speed and quicker delivery are other key factors looked by the online shoppers to choose their website/vender. Besides, few more critical factors like Secure and convenient payment method, Privacy, online support provided by the website are the deciding factors for online shoppers.

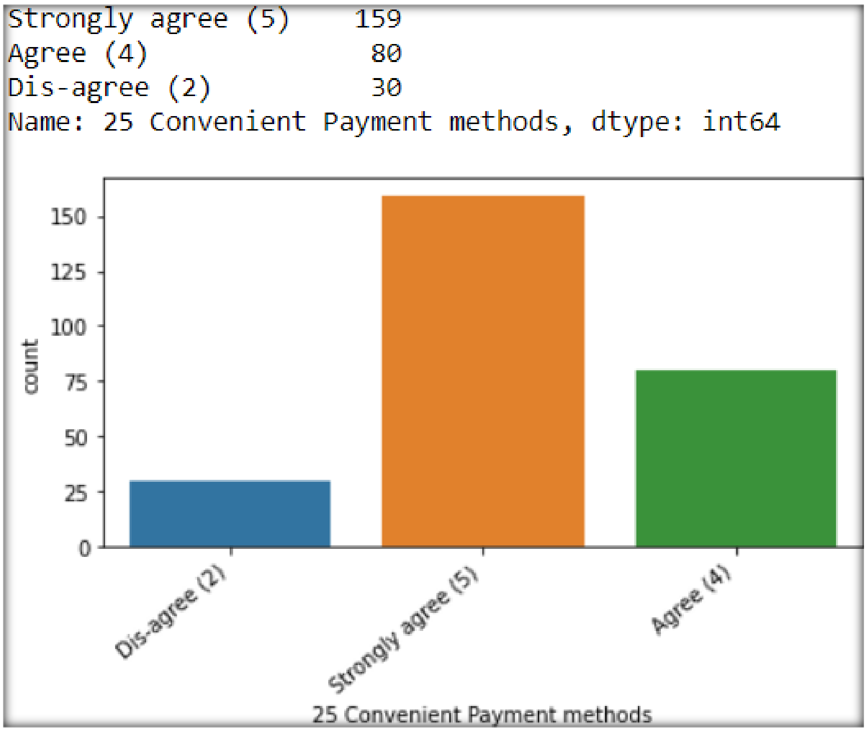
Visualizing such analysis would give an easy way to understanding. Preferred website selection visualization can give a clear picture.

If sites will be able to guarantee the privacy of the customer:



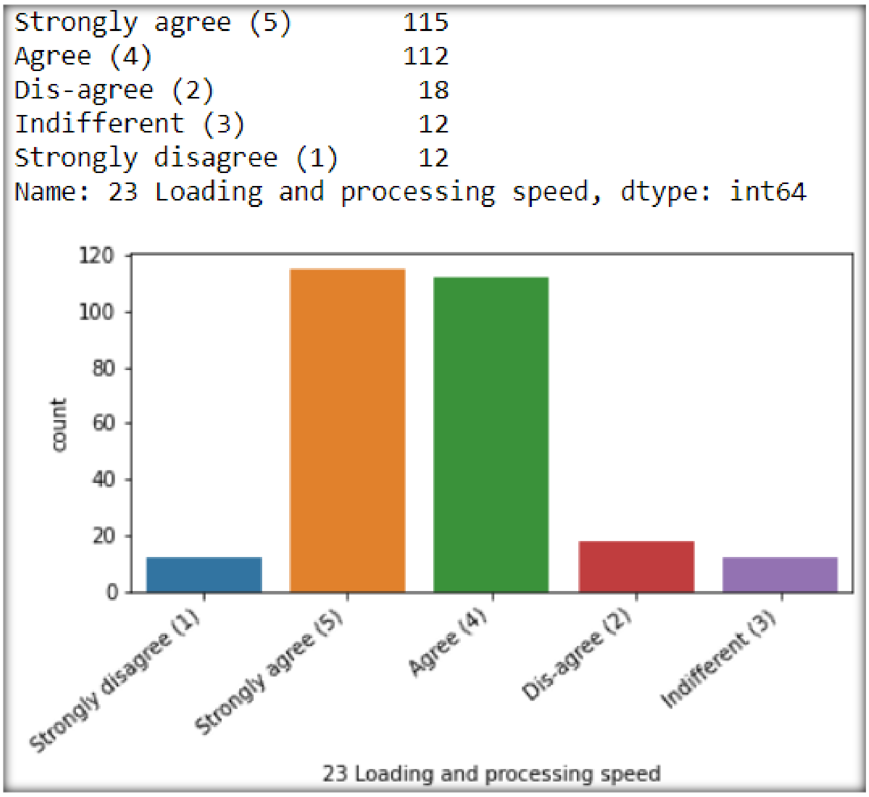
```
Strongly agree (5)    185
Agree (4)              58
indifferent (3)        26
Name: 28 Being able to guarantee the privacy of the customer
```

Secure and convenient payment options for users:



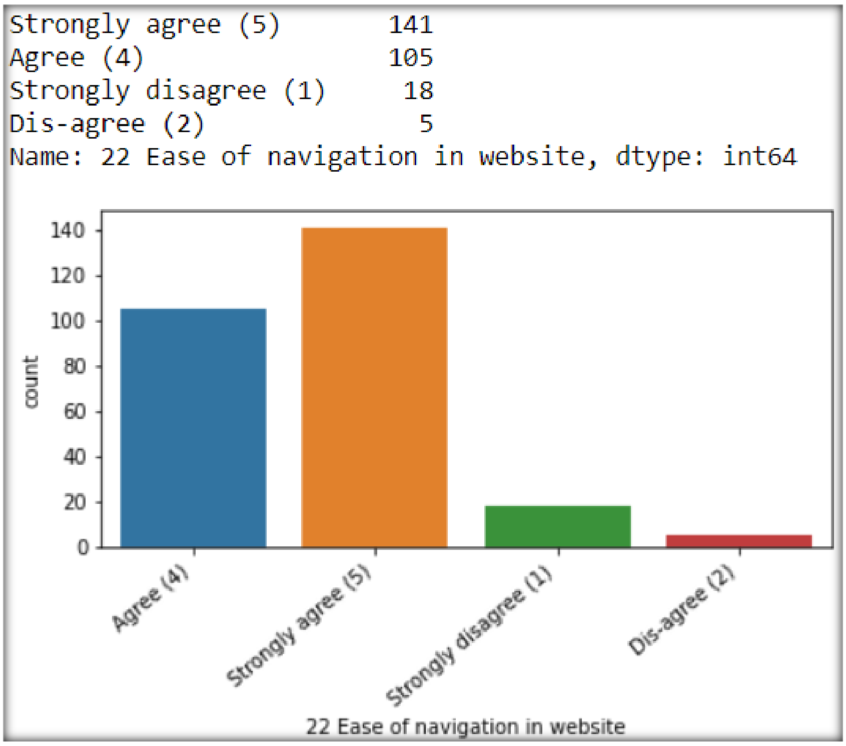
```
Strongly agree (5)    159
Agree (4)              80
Dis-agree (2)          30
Name: 25 Convenient Payment methods, dtype: int64
```

Performance of website impacts user’s preferences:



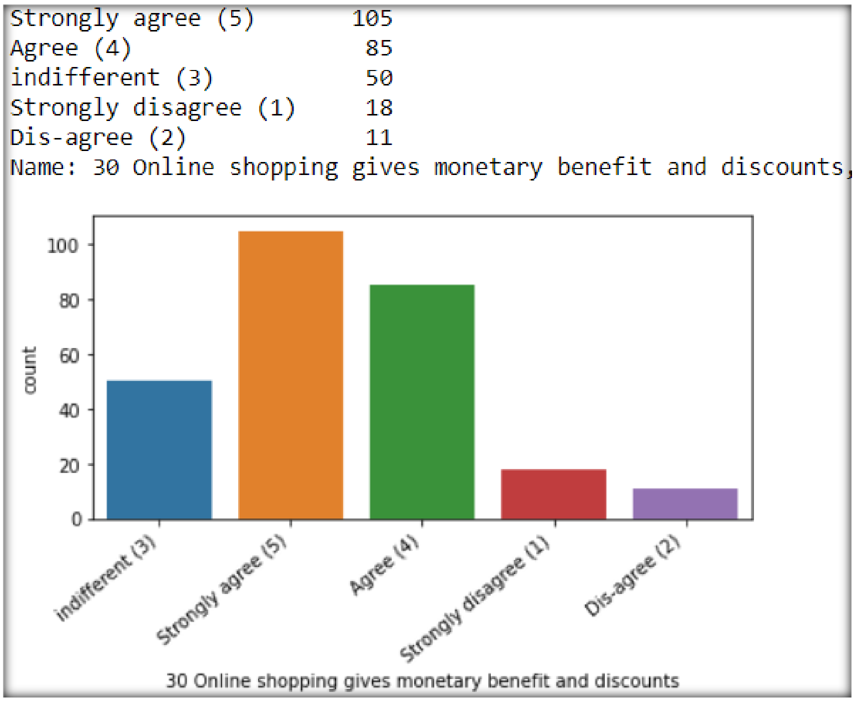
```
Strongly agree (5)       115
Agree (4)                112
Dis-agree (2)             18
Indifferent (3)           12
Strongly disagree (1)     12
Name: 23 Loading and processing speed, dtype: int64
```

The impact of easy and user-friendly navigations:



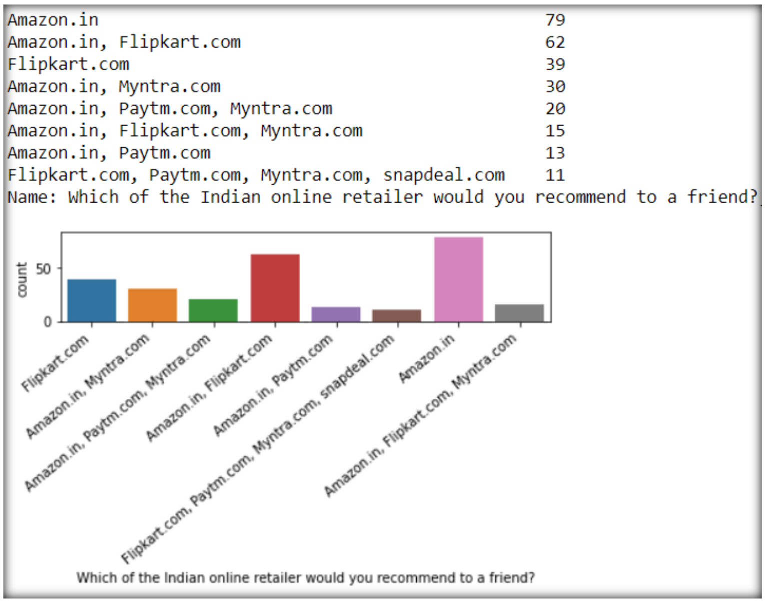
```
Strongly agree (5)       141
Agree (4)                105
Strongly disagree (1)     18
Dis-agree (2)              5
Name: 22 Ease of navigation in website, dtype: int64
```

Besides, discounts and locality bonus are the added advantages which an online shopper would be interested into.



```
Strongly agree (5)       105
Agree (4)                 85
indifferent (3)           50
Strongly disagree (1)     18
Dis-agree (2)             11
Name: 30 Online shopping gives monetary benefit and discounts,
```

With all of these and more analysis, we found that Amazon.in would be the preferred website/vender for online shopper in today's date.



```
Amazon.in                                         79
Amazon.in, Flipkart.com                           62
Flipkart.com                                      39
Amazon.in, Myntra.com                             30
Amazon.in, Paytm.com, Myntra.com                  20
Amazon.in, Flipkart.com, Myntra.com               15
Amazon.in, Paytm.com                              13
Flipkart.com, Paytm.com, Myntra.com, snapdeal.com    11
Name: Which of the Indian online retailer would you recommend to a friend?
```

count
50
0
Flipkart.com
Amazon.in, Myntra.com
Amazon.in, Paytm.com, Myntra.com
Amazon.in, Flipkart.com
Amazon.in, Paytm.com
Flipkart.com, Paytm.com, Myntra.com, snapdeal.com
Amazon.in
Amazon.in, Flipkart.com, Myntra.com
Which of the Indian online retailer would you recommend to a friend?

Model Building

a. Logistic Regression, where we got 68.11% accuracy.
b. Random Forest Classifier, where we got 100% accuracy.
c. Decision Tree Classifier, where we got 98.88% accuracy.
d. GaussianNB, where we got 84.04% accuracy.

Hardware and Software Requirements and Tools Used

a. Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram
b. Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

Identification of possible problem-solving approaches

Following models are used for solving the problem:

a. Logistic Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results).
b. Random Forest Classifier: is a meta estimator that fits several classifying decision tree on various sub samples of the dataset.
c. Decision Tree Classifier: A decision tree is a flowchart-like tree structure in which the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf

node represents the outcome. A Decision Tree consists of Nodes: Test for the value of a certain attribute.

d. Gaussian NB: Gaussian NB is based on the Naive Bayes theorem with the assumption of conditional independence between every pair of features given the label of the target class.
e. Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
f. Datasets: Overall dat.
g. Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.

Following statistical and analytical approach followed:

a. Started with univariate descriptive and graphs.
b. bivariate descriptive, again including graphs.
c. Model building and interpreting results.

Testing of Identified Approaches (Algorithms)

a. Train Test Split
b. Logistic Regression
c. Random Forest Classifier
d. Decision Tree Classifier
e. GaussianNB
f. Cross Validation
g. Hyper Parameter Tuning Using Grid Search CV.

Key Metrics for success in solving problem under consideration.

1. Cross Validation for cross validates the accuracy-score from overfitting.
2. Hyper parameter tuning using Grid Search Cv for making the prediction better.

Deployment

- All the models are performing well after cross validation Hyper parameter Tuning .
- We can use Decision Tree Classifier for our model.

References

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.

Learning Outcomes of the Study in respect of Data Science

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.