CUSTOMER RETENTION DATASET ANALYSIS

Submitted by:

ASHOK KUMAR SHARMA

ACKNOWLEDGMENT

I would like to express my deep sense of gratitude to my SME (Subject Matter Expert) **Mr. Shubham Yadav** as well as **Flip Robo Technologies** who gave me the golden opportunity to do this data analysis project on **Customer Retention Dataset Analysis**, which also helped me in doing lots of research and I came to know about so many new things.

I am very much thankful to **Dr. Deepika, Trainer (DataTrained)**, for their valuable guidance, keen interest and encouragement at various stages of my training period which eventually helped me a lot in doing this project.

I also acknowledge with thanks for suggestion and timely guidance, which I have received from my SME Mr. Shubham Yadav during this project, which immensely helped me in the evaluation of my ideas on the project.

ASHOK KUMAR SHARMA

INTRODUCTION

- ## Business Problem Framing

  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.

- ## Conceptual Background of the Domain Problem

  Customer retention refers to the ability of a company or product to retain its customers over some specified period. High customer

retention means customers of the product or business tend to return to, continue to buy or in some other way not defect to another product or business, or to non-use entirely. Selling organizations generally attempt to reduce customer defections. Customer retention starts with the first contact an organization has with a customer and continues throughout the entire lifetime of a relationship and successful retention efforts take this entire lifecycle into account. A company's ability to attract and retain new customers is related not only to its product or services, but also to the way it services its existing customers, the value the customers actually perceive as a result of utilizing the solutions, and the reputation it creates within and across the marketplace.

- Review of Literature

1. **What is Customer Retention?**

   Customer retention refers to the ability of a company or product to retain its customers over some specified period.

- Motivation for the Problem Undertaken

Successful customer retention involves more than giving the customer what they expect. Generating loyal advocates of the brand might mean exceeding customer expectations. Creating customer loyalty puts 'customer value rather than maximizing profits and shareholder value at the centre of business strategy'. The key differentiation in a competitive environment is often the delivery of a consistently high standard of customer service. Furthermore, in the emerging world of Customer Success, retention is a major objective.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  For checking datatypes and null values, padas.DataFrame.info() method has been used. To get the statistical summary overview, pandas.DataFrame.describe() method has been used to infer the following:

  1. Count: to count the number of records.
  2. Mean: to calculate the mean of the feature.
  3. Std: to calculate the Standard Deviation of the feature.
  4. Min: to find the minimum value of the feature.
  5. 25% (1^{st} Quartile): to find the first quartile of the feature.
  6. 50% (2^{nd} Quartile): to find the median or second quartile.
  7. 75% (3^{rd} Quartile): to find the third quartile of the feature.
  8. Max: to find the maximum value of the feature.

- Data Sources and their formats

  The dataset is being provided by Flib Robo Technologies in .xlsx (Microsoft Excel) format and contains 269 records with 71 features as explained below:

  1. **Gender 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?**
  10. **What is the operating system (OS) of your device?**
  11. **What browser do you run on your device to access the website?**
  12. **Which channel did you follow to arrive at your favorite online store for the first time?**
  13. **After first visit, how do you reach the online retail store?**

14. **How much time do you explore the e- retail store before making a purchase decision?**
15. **What is your preferred payment Option?**
16. **How 4 do you abandon (selecting an items and leaving without making payment) your shopping cart?**
17. **Why did you abandon the "Bag", "Shopping Cart"?**
18. **The content on the website must be easy to read and understand**
19. **Information on similar product to the one highlighted is important for product comparison**
20. **Complete information on listed seller and product being offered is important for purchase decision.**
21. **All relevant information on listed products must be stated clearly**
22. **Ease of navigation in website**
23. **Loading and processing speed**
24. **User friendly Interface of the website**
25. **Convenient Payment methods**
26. **Trust that the online retail store will fulfill its part of the transaction at the stipulated time**
27. **Empathy (readiness to assist with queries) towards the customers**
28. **Being able to guarantee the privacy of the customer**
29. **Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)**
30. **Online shopping gives monetary benefit and discounts**
31. **Enjoyment is derived from shopping online**
32. **Shopping online is convenient and flexible**
33. **Return and replacement policy of the e-tailer is important for purchase decision**
34. **Gaining access to loyalty programs is a benefit of shopping online**
35. **Displaying quality Information on the website improves satisfaction of customers**
36. **User derive satisfaction while shopping on a good quality website or application**
37. **Net Benefit derived from shopping online can lead to users satisfaction**
38. **User satisfaction cannot exist without trust**
39. **Offering a wide variety of listed product in several category**
40. **Provision of complete and relevant product information**
41. **Monetary savings**
42. **The Convenience of patronizing the online retailer**
43. **Shopping on the website gives you the sense of adventure**
44. **Shopping on your preferred e-tailer enhances your social status**
45. **You feel gratification shopping on your favorite e-tailer**
46. **Shopping on the website helps you fulfill certain roles**
47. **Getting value for money spent**
48. **From the following, tick any (or all) of the online retailers you have shopped from;**
49. **Easy to use website or application**
50. **Visual appealing web-page layout**
51. **Wild variety of product on offer**
52. **Complete, relevant description information of products**
53. **Fast loading website speed of website and application**
54. **Reliability of the website or application**
55. **Quickness to complete purchase**

56. **Availability of several payment options**
57. **Speedy order delivery**
58. **Privacy of customers' information**
59. **Security of customer financial information**
60. **Perceived Trustworthiness**
61. **Presence of online assistance through multi-channel**
62. **Longer time to get logged in (promotion, sales period)**
63. **Longer time in displaying graphics and photos (promotion, sales period)**
64. **Late declaration of price (promotion, sales period)**
65. **Longer page loading time (promotion, sales period)**
66. **Limited mode of payment on most products (promotion, sales period)**
67. **Longer delivery period**
68. **Change in website/Application design**
69. **Frequent disruption when moving from one page to another**
70. **Website is as efficient as before**
71. **Which of the Indian online retailer would you recommend to a friend?**

- Hardware and Software Requirements and Tools Used

  During this project following set of hardware is being used:

  RAM: 8 GB

  CPU: AMD A8 Quad Core 2.2 Ghz

  GPU: AMD Redon R5 Graphics

  and the following software and tools is being used:

  - Python
  - Jupyter Notebook
  - Anaconda

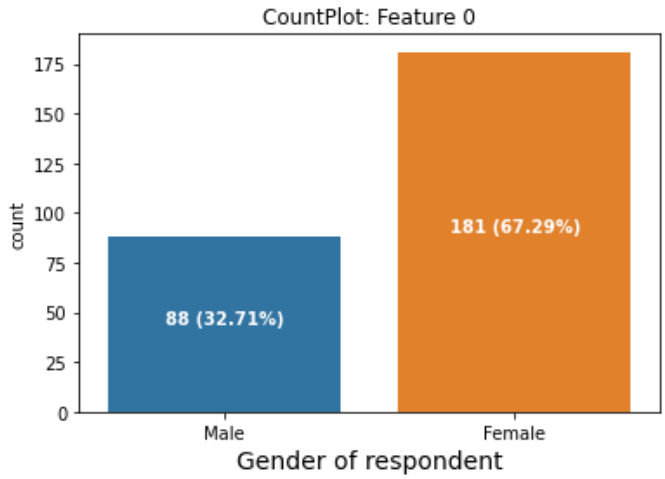
  With following libraries and packages:

  - Pandas
  - Matplotlib
  - Seaborn
  - sklearn

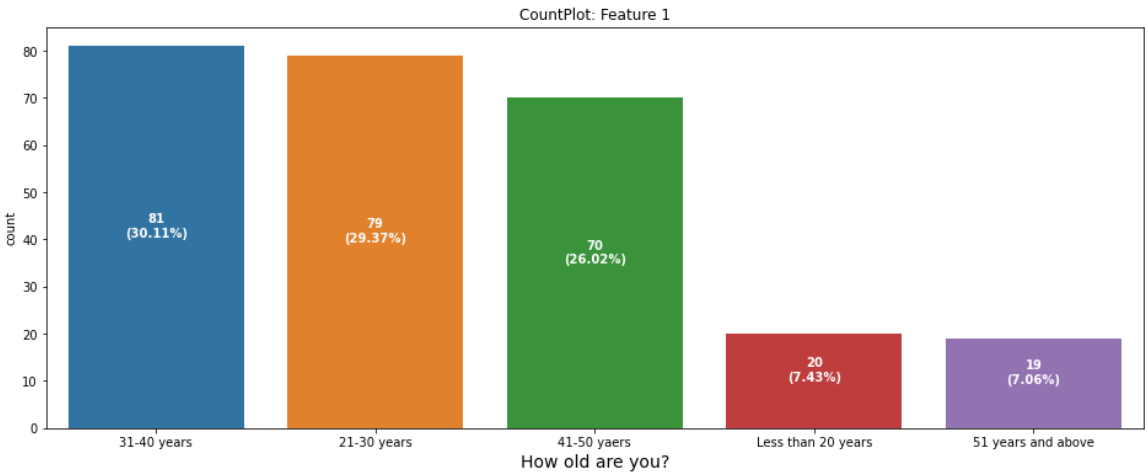
- Visualizations

  To better understand the data, following types of visualizations have been used: 1. Univariate, 2. Bivariate and 3. Multivariate.

  1. Univariate Analysis: Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. In this project count-plot has been used.



- From the above count plot of feature Gender of respondent it is clear that Most of the observations are for **Female with 67.29%** as compared to **Male with 32.71%** which indicates **dataset is imbalanced** and needs to be handled accordingly.



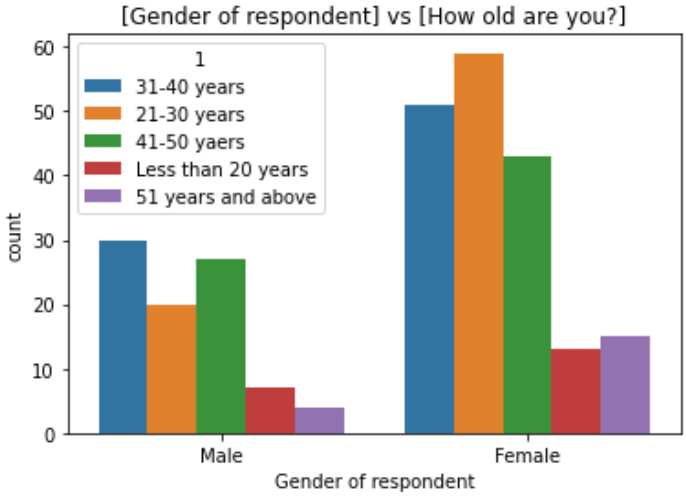
Remarks:

- Most of the records are for ages 21 to 50 years while less number of records are present for ages less than 20 years and greater than 50 years.

- Maximum number of records are present for ages 31-40 years while minimum number of records are present for ages 51 years and above.

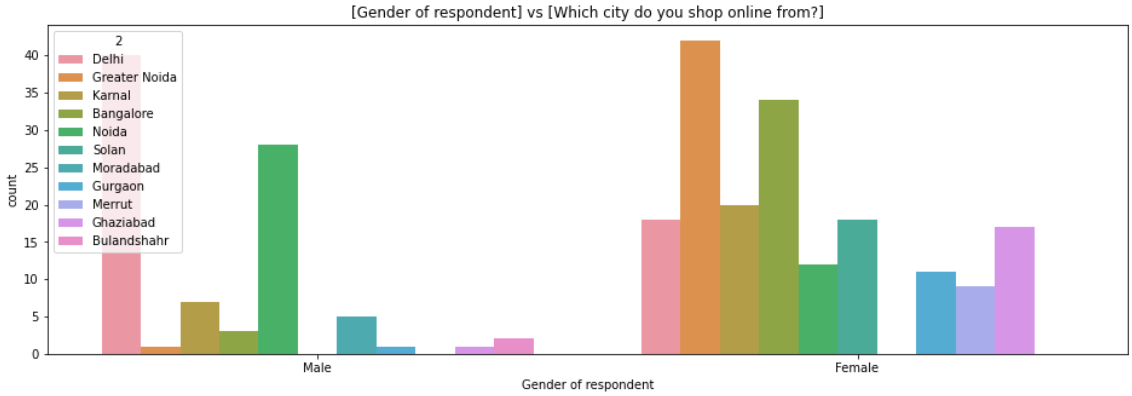
And so on…

2. Bivariate Analysis: Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them. We have analyzed the data and it's relationship with features using count plot as shown below:



Remarks:

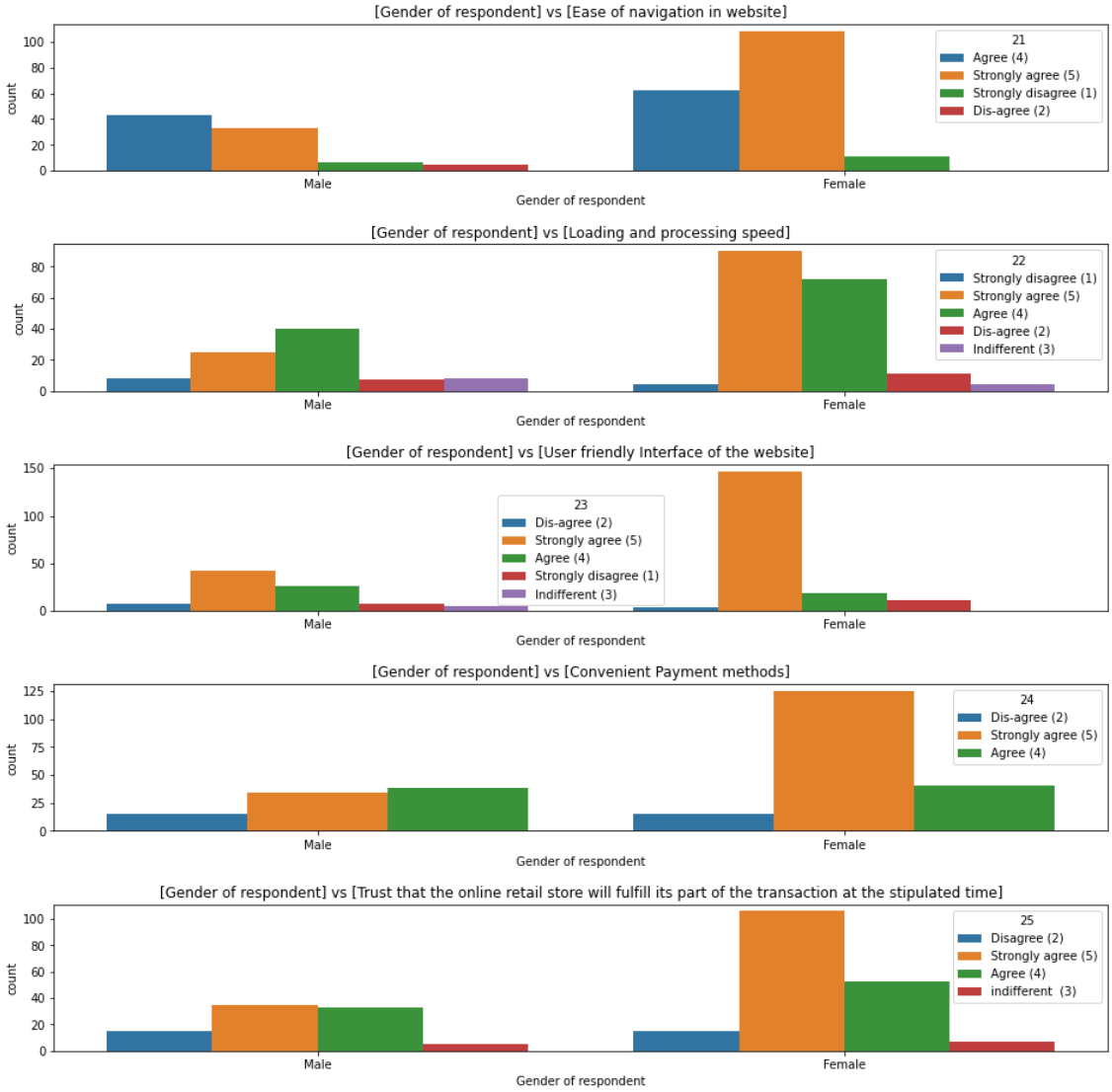
- Majority of Male and Female customers are of ages 21 to 50 years.
- Maximum number of Male customers ages from 31 to 40 years.
- Maximum number of Female customers ages from 21 to 30 years.
- Minimum number of Male and Female customers are of age less than 20 years.



Remarks:

- Majority of Male customers are from city `Delhi and Noida` whereas Female customers are from `Greater Noida and Banglore`.

- Maximum number of Male customers are form `Delhi` whereas Female customers are from `Greater Noida`.
- There are no Male customers from city `Solan and Merrut` whereas no Female customers are from `Moradabad and Bulandshahr`.



Remarks:

`for Feature 21: Ease of navigation in website`

- Most of the Male and Female customers Agree and Strongly agree.

`for Feature 22: Loading and processing speed`

- Most of the Male and Female customers Agree and Strongly agree.

`for Feature 23: User friendly Interface of the website`

- Most of the Male and Female customers Agree and Strongly agree.

`for Feature 24: Convenient Payment methods`

- Most of the Male and Female customers Agree and Strongly agree.

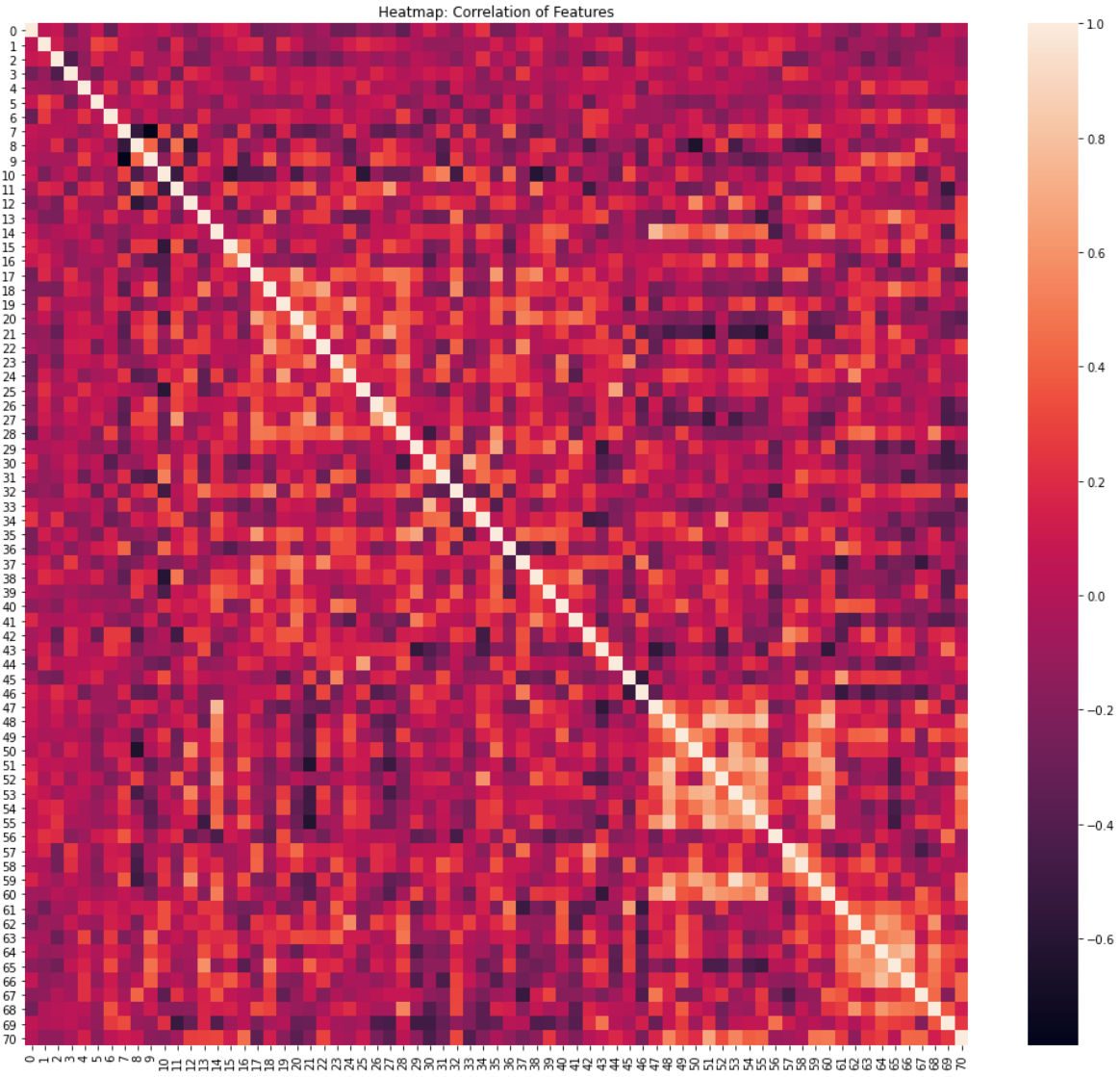
for Feature 25: Trust that the online retail store will fulfill its part of the transaction at the stipulated time

- Most of the Male and Female customers Agree and Strongly agree.

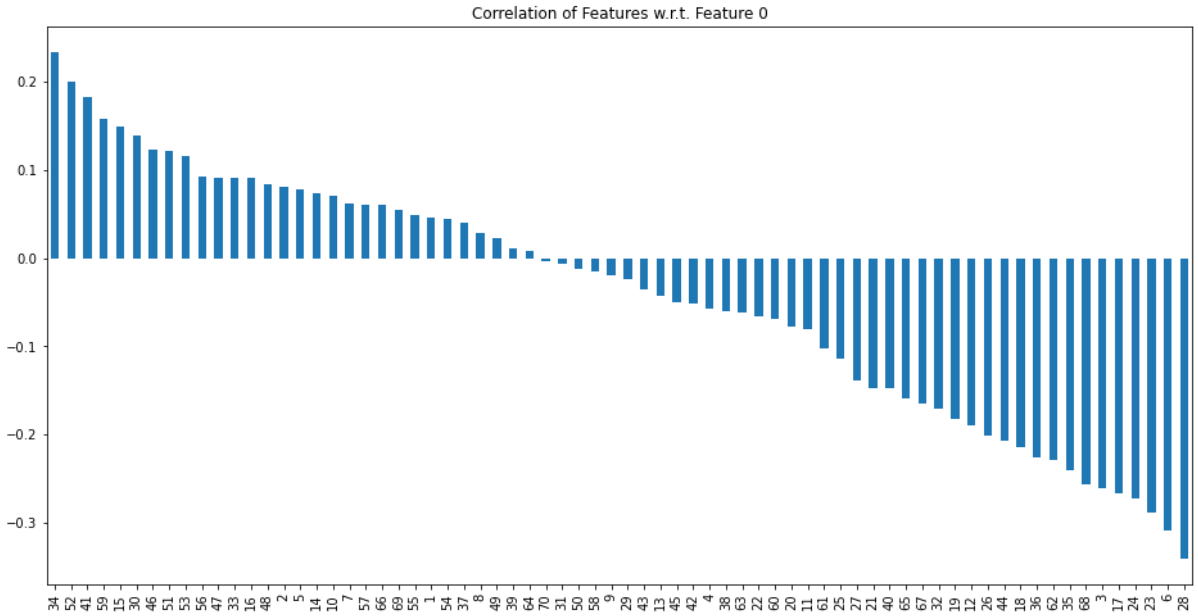
and so on…

3. Multivariate Analysis: Multivariate analysis is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.

- Heatmap is being used to represent the correlation of features from a scale of -1.0 to 1.0. After going through heatmap it is found that Feature 17 to 28, 47 to 60 and 61 to 68 are positive high correlated to each other while Feature 21 is negatively high correlated to feature 46 to 55.



For getting more accurate details regarding correlation of features w.r.t. Feature 0, we have used bar plot:



- With this plot it is clear that Feature 34, 52, 41, 59, 15, 30, 46, 51 and 53 shows good positive correlation with feature 0 while Feature 28, 6, 23, 24, 17, 3, 68,35, 62,36, 18, 44, 26, 12, 19 and 32 shows negatively good correlation with feature 0.

Interpretation of the Results / CONCLUSION

Starting with univariate analysis, we found that dataset is imbalanced by analysing Feature 0: Gender of respondent which consists of 32.71% records of Male and 67.29% of Female. This needs to be handled during the train test spilt part of model training. Moving further with count plot of features, we observed that most of the customers involved are of ages 21 to 50 years from city Delhi, Greater Noida, Karnal, Bangalore and Noida whereas very few involvements was found for city Moradabad, Merrut and Bulandshahr. We also found that for online shopping mostly mobile internet and Wi-Fi was used with Smartphone and Laptop having operating system Windows and Android and most of the time shopping website was accessed in Google Chrome browser. Also, most of the customers arrive to their favorite online store using Search Engine for the first time and then onwards they mostly use Search Engine, Via Application and Direct URL. We also found that most of customers takes more than 6 minutes to make a purchase decision at e-retail store and for the payment they mostly preferred Credit/Debit Cards and Cash on Delivery (CoD) option. Most of the customers abandon their Shopping Cart because of better alternative offer. We also found that most of customers looks for complete information regarding products, related products for comparison and seller. Most of the customers preferred websites which have ease of navigation, speedy loading, user friendly interface, payment convenient, secure transactions, complete privacy, prompt assistance, return & replacement policy and monitory savings. We also found that most of customers preferred Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com for online shoppings. After this we came up with bivariate analysis which gives us the close look of relationship between features. By using count plot with hue, we found that most of the male and female involved in online shoppings are of ages 21 to 50 years. Also, male in Delhi and Noida are more prominent to online shopping as compare to females while females in Greater Noida and Bangalore

are more prominent to online shoppings than males. We also found that use of Wi-Fi in online shopping is more by female as compare to male. After first visit to online retail store, female preferred to access website Via Application while in case of male, it is Search Engine and Direct URL. Also, it was found that female takes more time than male to make a purchase decision on an online store. On all other aspects of online shopping both male and female have same sort of views with certain differences. Moving further with multi-variate analysis, we found that feature 34, 52, 41, 59, 15, 30, 46, 51 and 53 shows good positive correlation with feature 0 while feature 28, 6, 23, 24, 17, 3, 68,35, 62,36, 18, 44, 26, 12, 19 and 32 shows negatively good correlation with feature 0.