E-Retail Customer Activation and Retention

Submitted by:

Ashish Kumar Samal

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Khusboo Garg as well as Flip Robo Technologies who gave me the opportunity to do this project on **'E-Retail Customer Activation and Retention'** & also helping me to gain in-depth knowledge of Machine Learning and DataScience to derive insights for organizational goals or meet business needs.

Also, I have utilized a few external resources that helped me to complete this project. All the external resources that were used in creating this project are listed below:

https://stackoverflow.com/questions

https://medium.com/

https://www.kaggle.com/

https://www.geeksforgeeks.org/

https://www.codegrepper.com/

https://www.analyticsvidhya.com/

https://towardsdatascience.com/

https://github.com/

INTRODUCTION

Business Problem Framing

Problem Overview

E-retail Customer Activation and Retention

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

Prediction:

We have to predict E-retail factors for customer activation and retention

Conceptual Background of the Domain Problem

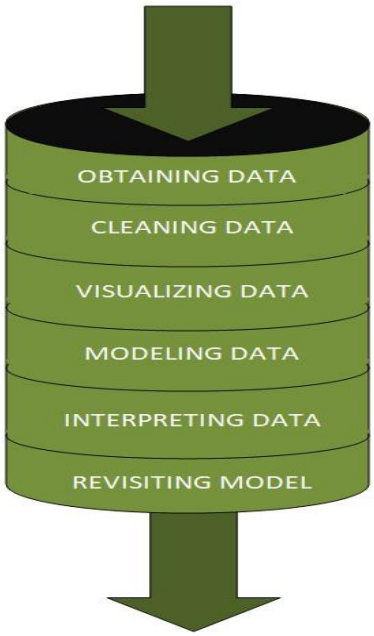
MACHINE LEARNING AND DATA SCIENCE FOR BUSINESS:

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn from experience, make predictions and gradually improving its accuracy. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data science will increase, requires to assist in the identification of the most relevant business questions and subsequently the data to answer them. Following are the ways Data science can add value to Business :

- Empowering management and officers to make better decision
- Directing actions based on trends—which in turn help to define goals
- Challenging the staff to adopt best practices and focus on issues that matter
- Identifying opportunities
- Decision making with quantifiable, data-driven evidence
- Testing these decisions
- Identification and refining of target audiences

DATASCIENCE PIPELINE:

The data science pipeline is a collection of connected tasks that aims at delivering an insightful data science product or service to the business organization. The responsibilities include collecting, cleaning, exploring, modeling, interpreting the data, and other processes of the launching of the product. This final product can be used for to achieve Business Goals.



Exploratory Data Analysis:

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

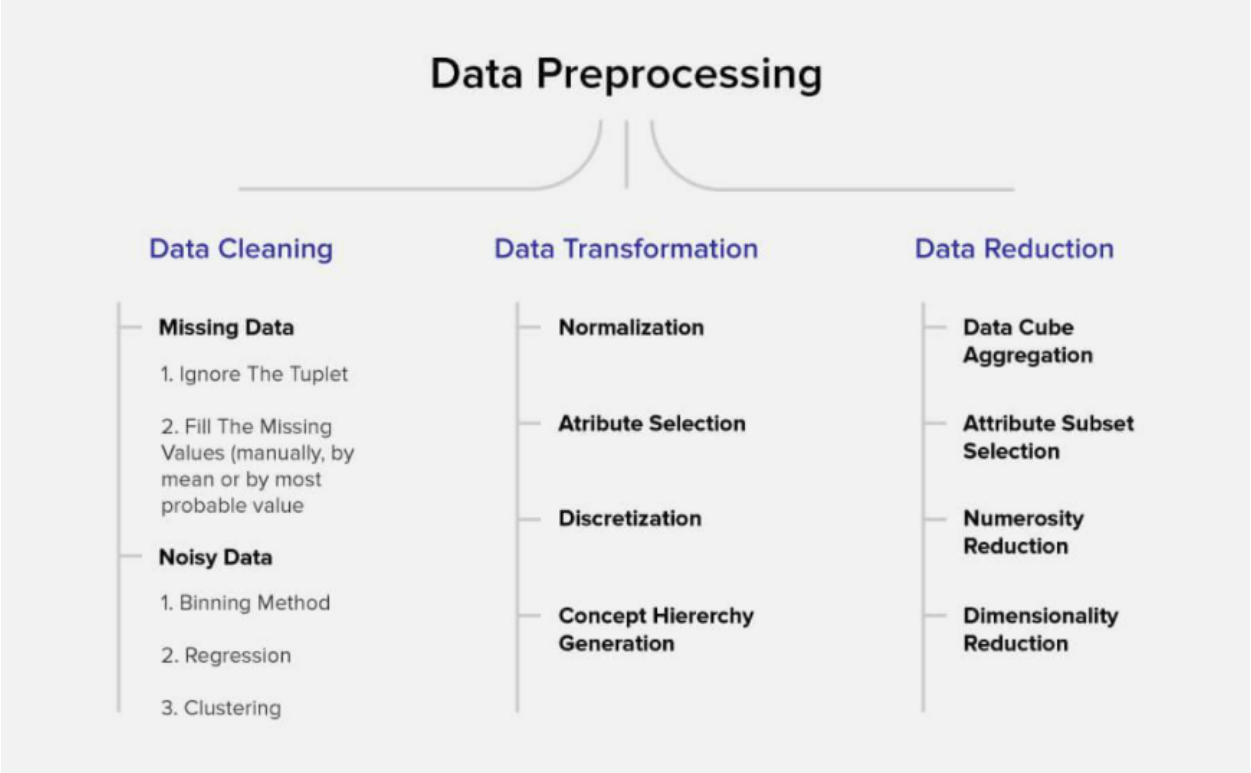
Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions

TYPES OF EXPLORATORY DATA ANALYSIS:

- Univariate Non-graphical
- Multivariate Non-graphical
- Univariate graphical
- Multivariate graphical

DATA PRE-PROCESSING & FEATURE ENGINEERING:

Preprocessing simply refers to perform series of operations to transform or change data. It is transformation applied to our data before feeding it to algorithm. When creating a machine learning project, and doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.



Data pre-processing is a very vital input to machine learning models, It is to prepare the raw data & make it suitable for efficient machine learning model. These are the methods of data preprocessing and we are going to use the required ones in our project.

FEATURE ENGINEERING:

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. As you may know, a “feature” is any measurable input that can be used in a predictive model.
Feature engineering, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

Feature Engineering Techniques for Machine Learning

- Imputation
- Handling Outliers
- Log Transform
- One-hot encoding/Label Encoding
- Scaling

Data Transformation:

Label Encoding:

As we mentioned above in library installation, Label Encoder is used to encode labels by assigning them numbers. It is used to encode single or multiple columns. Thus, if the feature is color with values such as ['white', 'red', 'black', 'blue']., using Label Encoder may encode color string label as [0, 1, 2, 3]

Handling Outliers:

The most important phase in Feature Engineering is handling outliers because it ensures that our model is trained on accurate data which leads to accurate models. An outlier may occur due to the variability in the data. It may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution). We have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

Below are some of the techniques of detecting outliers

- Boxplots
- Z-score

Variance Inflation Factor (VIF)

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library, Let us import the package.

SKEWNESS REMOVAL-(POWER-TRANSFORM):

Key step prior to initiating Machine learning models, optimizing, scaling the data to provide it as a input to start the modelling.

A power transform will make the probability distribution of a variable more Gaussian. This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. The log transform is a specific example of a family of transformations known as power transforms. The power_transform library present in the Sklearn. Pre-processing package.

MINMAX SCALER:

MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.
Before scaling we have to train test split the data.since we have to do skewness removal and scaling only on input data.

TRAIN TEST SPLIT:

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train_test_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train_test_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

Review of Literature

ABSTRACT:

E-retail has become the need of the hour for the modern customers nowadays. This project focuses on the key factors for customer activation & retention. Given dataset needs to be analysed in order to understand the things gone wrong / right to formulate a strategy / layout key points towards the customer activation & retention/

Motivation for the Problem Undertaken

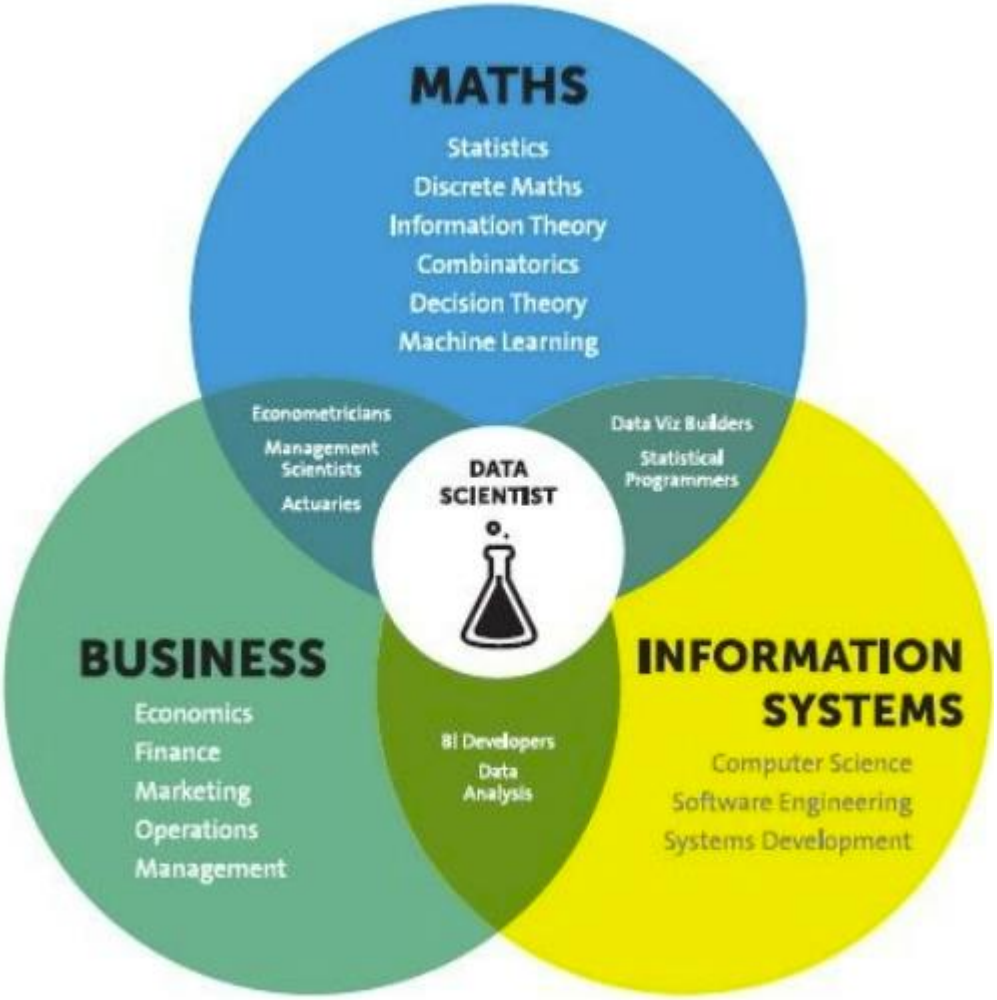
Business Goal:

We are required to analyse the India e-retail industry survey response dataset by bifurcating them into hedonic & Utilitarian values. Understand various influential factors customer encounters during online purchase done on a e-retail platform. Understand the customer's choice across various platform & layout the key indices which makes them to make repeated purchases on an certain e-retail platform.

Analytical Problem Framing

Mathematical/ Statistical /Analytical Modeling of the Problem

Mathematics, Statistics and Analytics are three of the most important concepts of Data Science. Data Science revolves around these three fields and draws their concepts to operate on the data.we will explore its practical usages in this field. So let's first explore how much these three are required for data science.



Mathematical Modelling

Mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization. Machine Learning is a field that focuses on computers having the ability to learn/operate without being programmed to do so.

Mathematics is playing an essential role in the latest technologies like Machine Learning, Artificial Intelligence, Data Science and Deep Learning, etc., It is because every algorithm built in the latest technologies has a mathematical function behind it and aid in identifying patterns.

The understanding of various notions of Statistics and Probability Theory are key for the implementation of such algorithms in data science. Notions include: Regression, Maximum Likelihood Estimation, the understanding of distributions (Binomial, Bernoulli, Gaussian (Normal)) and Bayes' Theorem.

The main reason for a greater significance of mathematics is because of its various concepts like: –

· Linear Algebra

· Probability

· Calculus

· Statistics

Linear Algebra & Calculus

Deep learning requires us to understand linear algebra & calculus, to understand how it works, for example forward propagation, backward propagation, parameters setting etc. For linear algebra, there are matrix operations (plus, minus, times, divide), scalar product, dot product, eigen-vectors and eigenvalues.

It is a branch of Mathematics for studying systems of equations. it can be one, two, and multi-dimensional equations. it helps us to solve numerical data or relations between two or more variables by establishing relations or equations between them. for example,

here' one basic algebraic equation:

$$y = a + bx + cx2$$

linear-algebra has a wide range of applications such as statics and matrices calculations, linear regression equations, descriptive statistics, graphic image vectors, Fourier series, graphs, and network establishment.

machine-learning algorithms like linear regression, logistic regression uses linear algebra to solve our target variables with given inputs/attributes or feature vectors given in the data set.

Calculus

Calculus is used essentially in optimization techniques. Using calculus, you can carry out mathematical modeling of artificial neural networks and also increase their accuracy and performance. For calculus, the data scientist need to understand various differentiation (to second-order derivative), integration, partial differentiation.

Differential Calculus

Differential Calculus studies the rate at which the quantities change. Derivates are most widely used for finding the maxima and minima of the functions. Derivates are used in optimization techniques where we have to find the minima in order to minimize the error function.

Integral Calculus

It is the mathematical study of the accumulation of quantities and for finding the area under the curve. Integrals are further divided into definite integrals and indefinite integrals.

Probability

The probability theory is very much helpful for making the prediction and Estimation.With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.

Probability is a very important mathematical concept for data science, used in validating hypothesis, bayes theorem and interpreting outputs in machine learning.

Bases on these we try to estimate various events, and the likelihood of the outcome. sometimes we wat graphical representations of probable outcomes which we call probability density functions or density curves.

Concepts of probability help us estimate expected value from given variables, to solve confusion matrix in classification algorithms, information entropy, evidence of particular attributes in naive Bayes classification, and even in statistics for hypothesis testings.

Statistics

A statistical model is a mathematical representation (or mathematical model) of observed data. When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.

So the areas in statistics are simple statistics like measurement of centrality, distributions and different probability distributions (Weibull, Poisson etc), Baye's Theorem

statistics is divided into two –

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

Descriptive Statistics or summary statistics is used for describing the data. It deals with the quantitative summarization of data. This summarization is performed through graphs or numerical representations.

Descriptive Statistics:

1) Mean, Median, Mode

2) IQR, percentiles

3) Std deviation and Variance

4) Normal Distribution

5) Z-statistics and T-statistics

6) correlation and linear regression

Inferential Statistics

It is the procedure of inferring or concluding from the data. Through inferential statistics, we make a conclusion about the larger population by running several tests and deductions from the smaller sample.

Inferential Statistics:

1) Sampling distributions

2) confidence interval

3) chi-square test

4) Advanced regression

5) ANOVA

The mathematical concepts noted above are key in understanding/implementing the following Machine Learning techniques.

- Supervised learning, including regression and classification models.

- Unsupervised learning, including clustering algorithms and association rules.

Regression Models

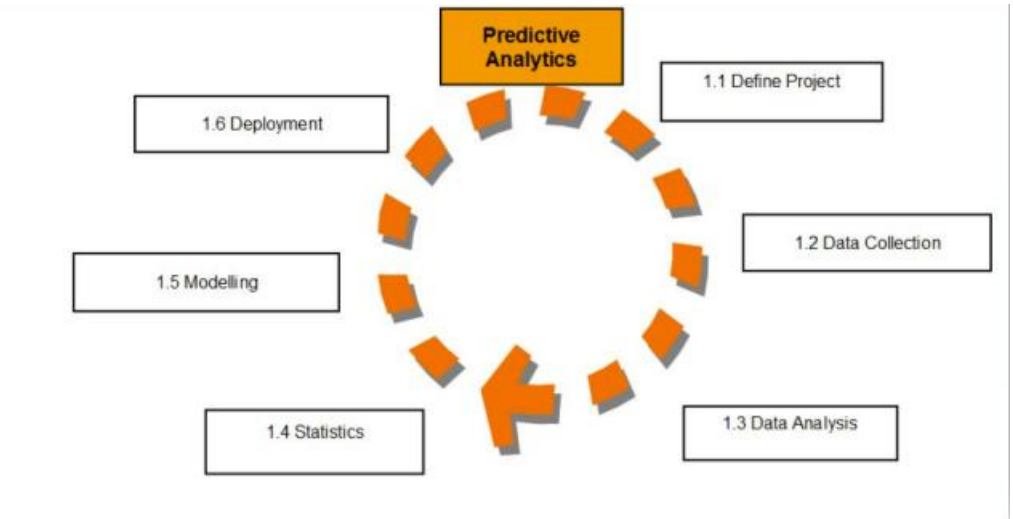
Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential business decisions.

Classification Models

Classification is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

Analytical Models:

An analytical model estimates or classifies data values by essentially drawing a line through data points. When applied to new data or records, a model can predict outcomes based on historical patterns.



. An analytical model is quantitative in nature, and used to answer a specific question or make a specific design decision. Different analytical models are used to address different aspects of the system, such as its performance, reliability, or mass properties.Data analysis comes with the fundamental types of data analytics encounter in data science: Descriptive, Diagnostic, Predictive, and Prescriptive.

- Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.
- Descriptive analysis is often used when reviewing any past or present data. This is because raw data is difficult to consume and interpret, while the metrics offered by descriptive analysis are much more focused.
- The example of descriptive statistics or analytics is to calculate the mean, median mode, standard deviation, and similar kinds of statistical calculation on finance or sales data.

- Diagnostic analytics takes it a step further to uncover the reasoning behind certain results. Diagnostic analytics is usually performed using such techniques as data discovery, drill-down, data mining, and different type of bivariant data analysis like correlations.etc.,
- Predictive Analytics is a **statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors**. Predictive Analytics can take both past and current data and offer predictions of what could happen in the future.
- Predictive models typically utilize variability in data to make the correct prediction and more variability of ingredient data that shows the relationship with what is possible to predict that united together into a prediction or valid score.
- Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, algorithms, and machine learning to make predictions and then suggests decision options to take advantage of the predictions. Prescriptive means (optimization and simulation).

Data Sources and their formats

Technical Requirements:

• Data contains 269 entries each having 71 variables.

• Data set doesn't contains Null values. We treated them using the domain knowledge and our own understanding.

• Extensive EDA has been performed to gain relationships of important variable and price.

• Data contains one numerical and all others as categorical variable. We handled them accordingly.

• We built Machine Learning models, applied regularization and determined the optimal values of Hyper Parameters.

• We found important features which affect the price positively or negatively.

The dataset is enclosed in notebook file

The dataset is provided to us by FlipRobo Technologies.And the dataset is in excel file format.

Data Description:

```
In [9]: df.columns

Out[9]: Index(['Gender of respondent', 'How old are you?',
               'Which city do you shop online from?',
               'What is the Pin Code of where you shop online from?',
               'Since How Long You are Shopping Online ?',
               'How many times you have made an online purchase in the past  year?',
               'How do you access the internet while shopping on-line?',
               'Which device do you use to access the online shopping?',
               'What is the screen size of your mobile device?',
               'What is the operating system (OS) of your device?',
               'What browser do you run on your device to access the website?',
               'Which channel did you follow to arrive at your favorite online store for the first time?',
               'After first visit, how do you reach the online retail store?',
               'How much time do you explore the e- retail store before making a purchase decision?',
               'What is your preferred payment Option?',
               'How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart?',
               'Why did you abandon the “Bag”, “Shopping Cart”?',
               'The content on the website must be easy to read and understand',
               'Information on similar product to the one highlighted  is important for product comparison',
               'Complete information on listed seller and product being offered is important for purchase decision.',
               'All relevant information on listed products must be stated clearly',
               'Ease of navigation in website', 'Loading and processing speed',
               'User friendly Interface of the website', 'Convenient Payment methods',
               'Trust that the online retail store will fulfill its part of the transaction at the stipulated time',
               'Empathy (readiness to assist with queries) towards the customers',
               'Being able to guarantee the privacy of the customer',
               'Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)',
               'Online shopping gives monetary benefit and discounts',
               'Enjoyment is derived from shopping online',
               'Shopping online is convenient and flexible',
               'Return and replacement policy of the e-tailer is important for purchase decision',
               'Gaining access to loyalty programs is a benefit of shopping online',
```

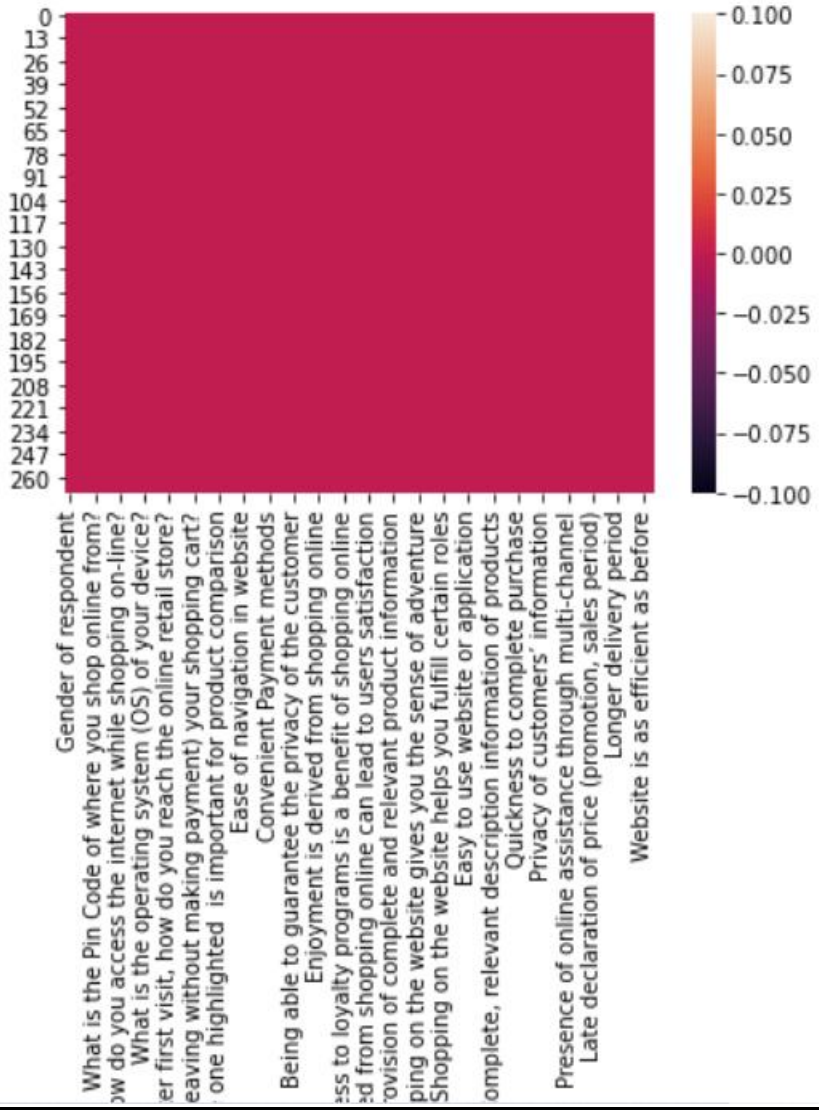
```
       'Displaying quality Information on the website improves satisfaction of customers',
       'User derive satisfaction while shopping on a good quality website or application',
       'Net Benefit derived from shopping online can lead to users satisfaction',
       'User satisfaction cannot exist without trust',
       'Offering a wide variety of listed product in several category',
       'Provision of complete and relevant product information',
       'Monetary savings',
       'The Convenience of patronizing the online retailer',
       'Shopping on the website gives you the sense of adventure',
       'Shopping on your preferred e-tailer enhances your social status',
       'You feel gratification shopping on your favorite e-tailer',
       'Shopping on the website helps you fulfill certain roles',
       'Getting value for money spent',
       'From the following, tick any (or all) of the online retailers you have shopped from;',
       'Easy to use website or application',
       'Visual appealing web-page layout', 'Wild variety of product on offer',
       'Complete, relevant description information of products',
       'Fast loading website speed of website and application',
       'Reliability of the website or application',
       'Quickness to complete purchase',
       'Availability of several payment options', 'Speedy order delivery',
       'Privacy of customers' information',
       'Security of customer financial information',
       'Perceived Trustworthiness',
       'Presence of online assistance through multi-channel',
       '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)',
       'Limited mode of payment on most products (promotion, sales period)',
       'Longer delivery period', 'Change in website/Application design',
       'Frequent disruption when moving from one page to another',
```

```
       'Website is as efficient as before',
       'Which of the Indian online retailer would you recommend to a friend?'],
      dtype='object')
```

Here the columns How old are you? ,How many times you have made an online purchase in the past 1 year?,What is the Pin Code of where you shop online from?,What is the Pin Code of where you shop online from?,What is the Pin Code of where you shop online from? are categorical ordinal data type.And all other columns are categorical nominal data type.Our Target column Which of the Indian online retailer would you recommend to a friend? is the categorical nominal data type.Hence it is a Classification Problem

```
sns.heatmap(df.isnull())
```

Out[15]: <AxesSubplot:>



There is no null values in the dataset.

DATA ACQUISITION

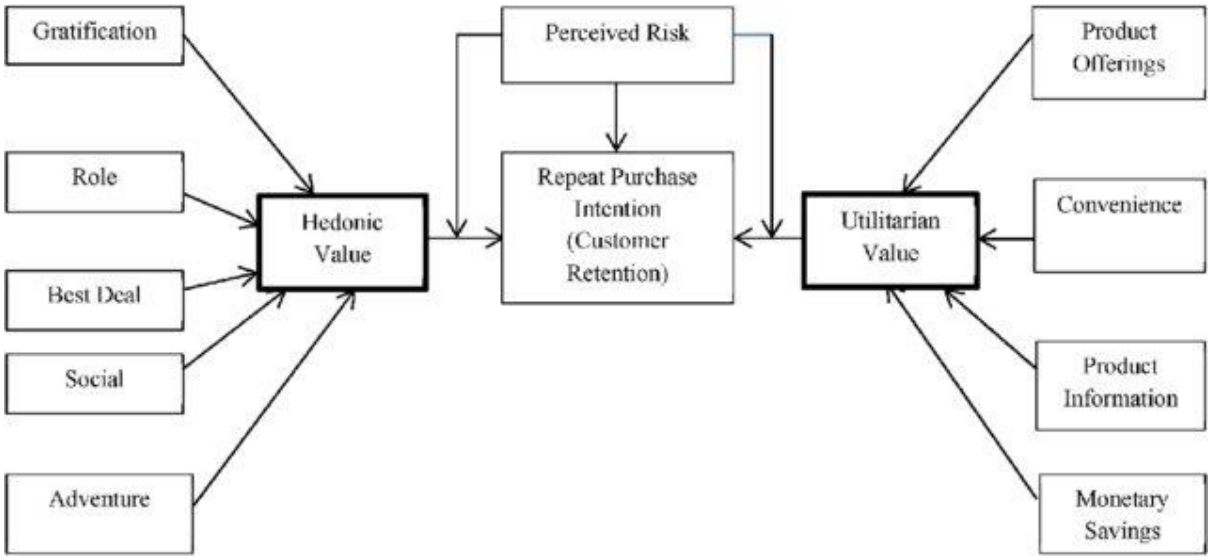
```
In [4]: #create dataframe
        df
```

	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\t	10 What is the operating system (OS) of your device? \t\t\t\t	browser do you run on your device to access the website? \t\t\t	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? \t\t\t\t	much time do you explore the e-retail store before making a purchase decision?
0	Male	31-40 years	Delhi	110009	Above 4 years	31-40 times	Dial-up	Desktop	Others	Window/windows Mobile	Google chrome	Search Engine	Search Engine	6-10 mins
1	Female	21-30 years	Delhi	110030	Above 4 years	41 times and above	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	Google chrome	Search Engine	Via application	more than 15 mins

FEATURE DESCRIPTION:

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

```
In [10]: df.dtypes

Out[10]: Gender of respondent                                          object
         How old are you?                                              object
         Which city do you shop online from?                           object
         What is the Pin Code of where you shop online from?            int64
         Since How Long You are Shopping Online ?                      object
                                                                        ...
         Longer delivery period                                        object
         Change in website/Application design                          object
         Frequent disruption when moving from one page to another      object
         Website is as efficient as before                             object
         Which of the Indian online retailer would you recommend to a friend?    object
         Length: 71, dtype: object
```



The left side are Hedonic values and the right side corner values are Utilitarian values.And these values we found in the dataset as attributes.These Attributes contributes as main impactful factors for the Customer Satisfaction and Retention.So our analysis also proved this with more further investigation.And also Found furthermore impactful factors for the E-Retail Business Enhancement.

Data Analysis and Preprocessing Done

```
In [14]: df.isnull().sum().any()
Out[14]: False
```

There are no null values in the dataset.

```
In [24]: positive=positive.rename(columns={0:'Positive suggestion',1:'Site',2:'Count'})
         positive
```

Out[24]:

	Positive suggestion	Site	Count
0	From the following, tick any (or all) of the o...	Amazon.in	269
1	From the following, tick any (or all) of the o...	Flipkart.com	221
2	From the following, tick any (or all) of the o...	Paytm.com	150
3	From the following, tick any (or all) of the o...	Myntra.com	146
4	From the following, tick any (or all) of the o...	Snapdeal.com	182
5	Easy to use website or application	Amazon.in	249
6	Easy to use website or application	Flipkart.com	201
7	Easy to use website or application	Paytm.com	125
8	Easy to use website or application	Myntra.com	147
9	Easy to use website or application	Snapdeal.com	130
10	Visual appealing web-page layout	Amazon.in	227

Remarks: From the above table, we can see the positive reviews with their counts.

```
In [28]: negative=negative.rename(columns={0:'Negative_review',1:'Site',2:'Count'})
         negative
```

Out[28]:

	Negative_review	Site	Count
0	Longer time to get logged in (promotion, sales...	Amazon.in	135
1	Longer time to get logged in (promotion, sales...	Flipkart.com	103
2	Longer time to get logged in (promotion, sales...	Paytm.com	77
3	Longer time to get logged in (promotion, sales...	Myntra.com	35
4	Longer time to get logged in (promotion, sales...	Snapdeal.com	67
5	Longer time in displaying graphics and photos ...	Amazon.in	126
6	Longer time in displaying graphics and photos ...	Flipkart.com	94
7	Longer time in displaying graphics and photos ...	Paytm.com	28
8	Longer time in displaying graphics and photos ...	Myntra.com	74
9	Longer time in displaying graphics and photos ...	Snapdeal.com	92
10	Late declaration of price (promotion, sales pe...	Amazon.in	56
11	Late declaration of price (promotion, sales pe...	Flipkart.com	43
12	Late declaration of price (promotion, sales pe...	Paytm.com	72
13	Late declaration of price (promotion, sales pe...	Myntra.com	75
14	Late declaration of price (promotion, sales pe...	Snapdeal.com	0
15	Longer page loading time (promotion, sales per...	Amazon.in	68

Remarks: From the above table, we can see the negative reviews with their count.

In [29]:
```
round(df['Which of the Indian online retailer would you recommend to a friend?'].value_counts(normalize=True)*100,2)
```

Out[29]:
```
Amazon.in                                          29.37
Amazon.in, Flipkart.com                            23.05
Flipkart.com                                       14.50
Amazon.in, Myntra.com                              11.15
Amazon.in, Paytm.com, Myntra.com                    7.43
Amazon.in, Flipkart.com, Myntra.com                 5.58
Amazon.in, Paytm.com                                4.83
Flipkart.com, Paytm.com, Myntra.com, snapdeal.com   4.09
Name: Which of the Indian online retailer would you recommend to a friend?, dtype: float64
```

we can see that the amazon.in and flipkart.com are the e-commerce sites that most customer prefers for shopping

seggregate the target column values to individual values because multiple labels present.We do this for analysis and model building purpose

In [30]:
```
df1=df.drop('Which of the Indian online retailer would you recommend to a friend?', axis=1).join(df['Which of the Indian online r
```

In [31]:
```
df1
```

Perceived stworthiness	Presence of online assistance through multi-channel	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)	Limited mode of payment on most products (promotion, sales period)	Longer delivery period	Change in website/Application design	Frequent disruption when moving from one page to another	Website is as efficient as before	Which of the Indian online retailer would you recommend to a friend?
Flipkart.com	Paytm.com	Amazon.in	Amazon.in	Flipkart.com	Flipkart.com	Amazon.in	Paytm.com	Flipkart.com	Amazon.in	Amazon.in	Flipkart.com
Myntra.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Flipkart.com	Myntra.com	snapdeal.com	Snapdeal.com	Snapdeal.com	Snapdeal.com	Amazon.in	Myntra.com	Amazon.in, Flipkart.com	Amazon.in

The target column values got segregated for analysis and prediction requirement

In [33]:
```
view=df[df['Which of the Indian online retailer would you recommend to a friend?']=='Flipkart.com']
view.loc[:,['Since How Long You are Shopping Online ?','How many times you have made an online purchase in the past  year?']]
```

Out[33]:

	Since How Long You are Shopping Online ?	How many times you have made an online purchase in the past year?
0	Above 4 years	31-40 times
9	Less than 1 year	Less than 10 times
23	Above 4 years	41 times and above
31	2-3 years	Less than 10 times
40	Above 4 years	31-40 times
44	Above 4 years	11-20 times
55	3-4 years	11-20 times
59	Above 4 years	41 times and above
66	3-4 years	Less than 10 times
71	3-4 years	Less than 10 times
85	Above 4 years	Less than 10 times
90	Less than 1 year	Less than 10 times
92	Above 4 years	31-40 times
98	Less than 1 year	Less than 10 times
106	1-2 years	31-40 times
113	Above 4 years	31-40 times

This result shows how long shopping made online and how many times purchase made in last year for flipkart

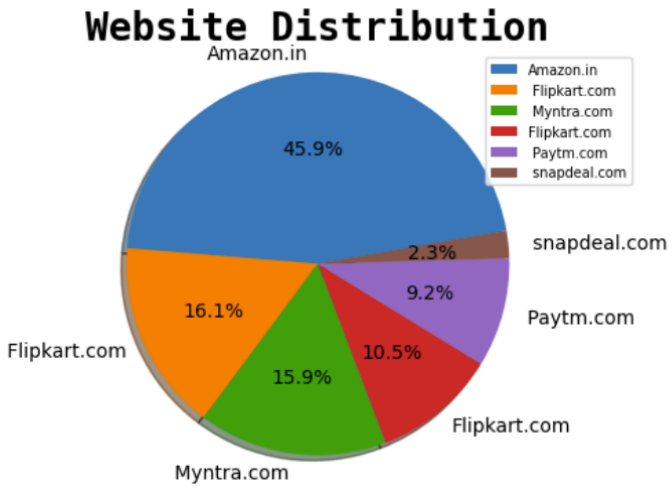
In [34]:
```
view1=df[df['Which of the Indian online retailer would you recommend to a friend?']=='Amazon.in']
view1.loc[:,['Since How Long You are Shopping Online ?','How many times you have made an online purchase in the past  year?']]
```

Out[34]:

	Since How Long You are Shopping Online ?	How many times you have made an online purchase in the past year?
7	3-4 years	Less than 10 times
8	2-3 years	Less than 10 times
10	Above 4 years	21-30 times
21	3-4 years	31-40 times
22	Above 4 years	31-40 times
24	3-4 years	41 times and above
38	Above 4 years	41 times and above
39	2-3 years	31-40 times
41	3-4 years	41 times and above
52	Less than 1 year	41 times and above
53	Above 4 years	41 times and above

This result shows how long shopping made online and how many times purchase made in last year for Amazon site

In [35]:
```
# Creating a pie chart
ax = df1['Which of the Indian online retailer would you recommend to a friend?'].value_counts()

plt.style.use('default')
plt.figure(figsize=(6, 4))
plt.pie(ax.values, labels=ax.index, startangle=10, explode=(None), shadow=True, autopct='%1.1f%%')
plt.title('Website Distribution', fontdict={
          'fontname': 'Monospace', 'fontsize': 20, 'fontweight': 'bold'})
plt.legend()
plt.legend(prop={'size': 7})
plt.axis('equal')
plt.show()
```



the above result shows shows amazon is the most customer friendly,activated and retention website among others

In [36]:
```
Amazon_usage=df1.groupby(['Which of the Indian online retailer would you recommend to a friend?','Since How Long You are Shopping
Amazon_usage
```

Out[36]:

	Which of the Indian online retailer would you recommend to a friend?	Since How Long You are Shopping Online ?	0
0	Amazon.in	Above 4 years	73
1	Amazon.in	2-3 years	63
2	Amazon.in	3-4 years	40
3	Flipkart.com	Above 4 years	34
4	Amazon.in	Less than 1 year	33
5	Flipkart.com	Above 4 years	25
6	Myntra.com	2-3 years	25
7	Myntra.com	Above 4 years	23
8	Paytm.com	Above 4 years	19
9	Flipkart.com	2-3 years	18
10	Flipkart.com	Less than 1 year	16
11	Myntra.com	3-4 years	13
12	Myntra.com	Less than 1 year	12
13	Paytm.com	3-4 years	11
14	Amazon.in	1-2 years	10
15	Flipkart.com	Less than 1 year	10

the above result shows amazn is the most shopped website among others for years

In [37]:
```
location_usage=df1.groupby(['Which of the Indian online retailer would you recommend to a friend?','What is your preferred paymer
location_usage
```

Out[37]:

	Which of the Indian online retailer would you recommend to a friend?	What is your preferred payment Option?	0
0	Amazon.in	Credit/Debit cards	137
1	Flipkart.com	Credit/Debit cards	70
2	Myntra.com	Credit/Debit cards	56
3	Amazon.in	Cash on delivery (CoD)	49
4	Paytm.com	E-wallets (Paytm, Freecharge etc.)	33
5	Amazon.in	E-wallets (Paytm, Freecharge etc.)	33
6	Flipkart.com	Cash on delivery (CoD)	27
7	Myntra.com	E-wallets (Paytm, Freecharge etc.)	20
8	Flipkart.com	E-wallets (Paytm, Freecharge etc.)	12
9	Paytm.com	Credit/Debit cards	11
10	snapdeal.com	Credit/Debit cards	11
11	Flipkart.com	Credit/Debit cards	11
12	Flipkart.com	Cash on delivery (CoD)	7

this shows that the amazon is the most shopping website by the customers and they use credit or debit card for the transaction among other payment methods

In [38]:
```
abandon_sites_count=df1.groupby(['Which of the Indian online retailer would you recommend to a friend?','Why did you abandon the
abandon_sites_count
```

Out[38]:

	Which of the Indian online retailer would you recommend to a friend?	Why did you abandon the “Bag”, “Shopping Cart”?	0
0	Amazon.in	Better alternative offer	133
1	Flipkart.com	Better alternative offer	77
2	Myntra.com	Promo code not applicable	46
3	Amazon.in	Promo code not applicable	43
4	Paytm.com	Promo code not applicable	31
5	Flipkart.com	Lack of trust	31
6	Myntra.com	Better alternative offer	30
7	Amazon.in	Change in price	29
8	Amazon.in	No preferred mode of payment	14
9	Paytm.com	Better alternative offer	13
10	snapdeal.com	Promo code not applicable	11
11	Flipkart.com	Promo code not applicable	11
12	Flipkart.com	Change in price	8

the customers abandon the purchase before checkout because there are more better alternative offer available.And also the case when the promo code not applicable most of the times

HEDONIC VALUES

In [39]:
```
hidonic_value=df1.groupby(['Which of the Indian online retailer would you recommend to a friend?','Shopping on the website gives
hidonic_value
```

Out[39]:

	Which of the Indian online retailer would you recommend to a friend?	Shopping on the website gives you the sense of adventure	Shopping on your preferred e-tailer enhances your social status	You feel gratification shopping on your favorite e-tailer	Shopping on the website helps you fulfill certain roles	Getting value for money spent	0
0	Flipkart.com	Agree (4)	Strongly agree (5)	Agree (4)	indifferent (3)	Agree (4)	25
1	Amazon.in	Agree (4)	Strongly agree (5)	Agree (4)	indifferent (3)	Agree (4)	25
2	Myntra.com	Agree (4)	Agree (4)	indifferent (3)	indifferent (3)	Agree (4)	20
3	Paytm.com	Agree (4)	Agree (4)	indifferent (3)	indifferent (3)	Agree (4)	20
4	Amazon.in	Agree (4)	Agree (4)	indifferent (3)	indifferent (3)	Agree (4)	20
5	Flipkart.com	Agree (4)	indifferent (3)	indifferent (3)	indifferent (3)	Agree (4)	19
6	Amazon.in	Dis-agree (2)	Agree (4)	Agree (4)	indifferent (3)	Agree (4)	19
7	Flipkart.com	Dis-agree (2)	Agree (4)	Agree (4)	indifferent (3)	Agree (4)	19
8	Amazon.in	Strongly agree (5)	Strongly disagree (1)	Strongly disagree (1)	Strongly disagree (1)	Agree (4)	18
9	Flipkart.com	Dis-agree (2)	indifferent (3)	indifferent (3)	Dis-agree (2)	Agree (4)	15
10	Amazon.in	indifferent (3)	indifferent (3)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	15
11	Myntra.com	Dis-agree (2)	indifferent (3)	indifferent (3)	Dis-agree (2)	Agree (4)	15

this shows customers acceptance over hedonic values provided by the respective websites.Amazon and Flipkart are doing well in customer retention with hedonic values among other websites

UTILITARIAN VALUE

In [40]:
```
utilitarian_value=df1.groupby(['Which of the Indian online retailer would you recommend to a friend?','Offering a wide variety of
utilitarian_value
```

Out[40]:

	Which of the Indian online retailer would you recommend to a friend?	Offering a wide variety of listed product in several category	Provision of complete and relevant product information	Monetary savings	The Convenience of patronizing the online retailer	0
0	Amazon.in	Agree (4)	Agree (4)	Strongly agree (5)	Agree (4)	40
1	Amazon.in	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	39
2	Flipkart.com	Agree (4)	Agree (4)	Strongly agree (5)	Agree (4)	25
3	Myntra.com	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Agree (4)	20
4	Paytm.com	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Agree (4)	20
5	Amazon.in	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Agree (4)	20
6	Amazon.in	indifferent (3)	Strongly agree (5)	Strongly agree (5)	indifferent (3)	19
7	Flipkart.com	indifferent (3)	Strongly agree (5)	Strongly agree (5)	indifferent (3)	19

this shows customers acceptance over utilitatian values provided by the respective websites.Amazon is doing well in customer retention with utilitatian values among other websites

In [41]:
```
ability of the website or application','Website is as efficient as before']).size().sort_values(0, ascending=False).reset_index()
```

Out[41]:

	Which of the Indian online retailer would you recommend to a friend?	Easy to use website or application	Wild variety of product on offer	Complete, relevant description information of products	Perceived Trustworthiness	Availability of several payment options	Reliability of the website or application	Website is as efficient as before	0
0	Flipkart.com	Amazon.in, Flipkart.com, Paytm.com, Myntra.com...	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Snapdeal.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Flipkart.com, Paytm.com	Amazon.in, Flipkart.com, Paytm.com	25
1	Amazon.in	Amazon.in, Flipkart.com, Paytm.com, Myntra.com...	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Snapdeal.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Flipkart.com, Paytm.com	Amazon.in, Flipkart.com, Paytm.com	25
2	Myntra.com	Amazon.in, Paytm.com, Myntra.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com, Myntra.com	Amazon.in, Myntra.com	Patym.com, Myntra.com	Amazon.in, Paytm.com, Myntra.com	Amazon.in	20
3	Amazon.in	Amazon.in, Paytm.com, Myntra.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com, Myntra.com	Amazon.in, Myntra.com	Patym.com, Myntra.com	Amazon.in, Paytm.com, Myntra.com	Amazon.in	20

The top most good customer retention websites are amazon and flipkart

In [42]:
```
customer_opinion_for_retention=df1.groupby(['Which of the Indian online retailer would you recommend to a friend?','User satisfac
customer_opinion_for_retention
```

Out[42]:

	Which of the Indian online retailer would you recommend to a friend?	User satisfaction cannot exist without trust	Information on similar product to the one highlighted is important for product comparison	All relevant information on listed products must be stated clearly	Convenient Payment methods	Gaining access to loyalty programs is a benefit of shopping online	Online shopping gives monetary benefit and discounts	0
0	Amazon.in	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	36
1	Flipkart.com	Agree (4)	Agree (4)	Agree (4)	Strongly agree (5)	indifferent (3)	Agree (4)	25
2	Amazon.in	Agree (4)	Agree (4)	Agree (4)	Strongly agree (5)	indifferent (3)	Agree (4)	25
3	Amazon.in	Agree (4)	Agree (4)	Agree (4)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	20
4	Myntra.com	Agree (4)	Agree (4)	Agree (4)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	20
5	Paytm.com	Agree (4)	Agree (4)	Agree (4)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	20
6	Flipkart.com	Agree (4)	Strongly agree (5)	Strongly agree (5)	Agree (4)	Agree (4)	indifferent (3)	19
7	Amazon.in	Agree (4)	Strongly agree (5)	Strongly agree (5)	Agree (4)	Agree (4)	indifferent (3)	19
8	Flipkart.com	Agree (4)	Strongly agree (5)	Agree (4)	Agree (4)	Agree (4)	Agree (4)	19
9	Amazon.in	Strongly disagree (1)	Dis-agree (2)	Strongly disagree (1)	Dis-agree (2)	Strongly agree (5)	Strongly disagree (1)	18
10	Amazon.in	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Strongly agree (5)	Agree (4)	15

All the utilitarian values are provided mostly by amazon only when compared with others

Since we have no null values and we segregated the target column.so data is clean now.

Data Inputs- Logic- Output Relationships

In Classification, the output variable must be **a discrete value**.

In classification, **inputs are divided into two or more classes**, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way.

In machine learning, classification refers to **a predictive modeling problem where a class label is predicted for a given example of input data**. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. The output variables are often called **labels or categories**. ... A classification can have real-valued or discrete input variables. A problem with two classes is often called a two-class or binary classification problem. A problem with more than two classes is often called a multi-class classification problem. A classification algorithm, in general, is a function that weighs the input features so that the output **separates one class into positive values and the other into negative values**. Classification is a data mining function that **assigns items in a collection to target categories or classes**. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

Classification analysis is a data analysis task within data-mining, **that identifies and assigns categories to a collection of data to allow for more accurate analysis**.

... Classification analysis can be used to question, make a decision, or predict behavior through the use of an algorithm.

DATA PREPROCESSING AND FEATURE ENGINEERING

DATA PRE-PROCESSING And FEATURE ENGINEERING

In [58]:
```
#converting string data type to int type using LabelEncoding
le=LabelEncoder()

list1=['Gender of respondent', 'How old are you?',
       'Which city do you shop online from?',
       'Since How Long You are Shopping Online ?',
       'How many times you have made an online purchase in the past  year?',
       'How do you access the internet while shopping on-line?',
       'Which device do you use to access the online shopping?',
       'What is the screen size of your mobile device?',
       'What is the operating system (OS) of your device?',
       'What browser do you run on your device to access the website?',
       'Which channel did you follow to arrive at your favorite online store for the first time?',
       'After first visit, how do you reach the online retail store?',
       'How much time do you explore the e- retail store before making a purchase decision?',
       'What is your preferred payment Option?',
       'How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart?',
       'Why did you abandon the “Bag”, “Shopping Cart”?',
       'The content on the website must be easy to read and understand',
       'Information on similar product to the one highlighted  is important for product comparison',
       'Complete information on listed seller and product being offered is important for purchase decision.',
       'All relevant information on listed products must be stated clearly',
       'Ease of navigation in website', 'Loading and processing speed',
       'User friendly Interface of the website', 'Convenient Payment methods',
       'Trust that the online retail store will fulfill its part of the transaction at the stipulated time',
       'Empathy (readiness to assist with queries) towards the customers',
       'Being able to guarantee the privacy of the customer',
       'Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)',
       'Frequent disruption when moving from one page to another',
       'Website is as efficient as before',
       'Which of the Indian online retailer would you recommend to a friend?']
for val in list1:
    df1[val]=le.fit_transform(df1[val].astype(str))
df1
```

	Gender of respondent	How old are you?	Which city do you shop online from?	What is the Pin Code of where you shop online from?	Since How Long You are Shopping Online ?	How many times you have made an online purchase in the past year?	How do you access the internet while shopping on-line?	Which device do you use to access the online shopping?	What is the screen size of your mobile device?	What is the operating system (OS) of your device?	What browser do you run on your device to access the website?	Which channel did you follow to arrive at your favorite online store for the first time?	After first visit, how do you reach the online retail store?	How much time do you explore the e-retail store before making a purchase decision?	What is your preferred payment Option?	fr
0	1	1	2	110009	3	2	0	0	3	2	0	2	2	2	2	
1	0	0	2	110030	3	3	3	2	0	1	0	2	4	4	1	
1	0	0	2	110030	3	3	3	2	0	1	0	2	4	4	1	
2	0	0	4	201308	2	3	1	2	2	0	0	2	4	1	2	
2	0	0	4	201308	2	3	1	2	2	0	0	2	4	1	2	

```
In [59]: #find correlation co.efficient of all variables in table
         df1.corr()
```

864	0.525202	1.000000	0.419304	-0.071498	0.040939	0.026372	0.260573	-0.200472	0.204252	-0.075731	-0.153756	0.002398	0.000767
310	0.346688	0.419304	1.000000	0.292234	0.070081	-0.071330	0.020699	-0.289108	-0.132036	0.513085	0.248801	-0.221852	-0.110405
345	0.052109	-0.071498	0.292234	1.000000	0.467938	0.190173	0.363089	0.238982	0.044900	0.262896	0.216116	-0.079639	-0.271973
069	0.172626	0.040939	0.070081	0.467938	1.000000	0.484187	0.627594	0.511901	0.253313	-0.048516	0.557905	-0.144191	-0.369901
848	0.202559	0.026372	-0.071330	0.190173	0.484187	1.000000	0.559522	0.571000	0.609806	-0.107399	0.424463	0.150679	-0.298403

```
In [63]: # Variables more than 0.70 correlations
         c = df1.corr().abs()
         s = c.unstack()
         so = s.sort_values(kind="quicksort", ascending=False)
         df_corr = pd.DataFrame(so)
         #df_corr.columns = ['correlations']
         print(df_corr[(df_corr[0] < 1) & (df_corr[0] > 0.7)])
```

		0
Reliability of the website or application	Perceived Trustworthiness	0.923353
Perceived Trustworthiness	Reliability of the website or application	0.923353
Easy to use website or application	Availability of several payment options	0.841864
Availability of several payment options	Easy to use website or application	0.841864
Easy to use website or application	Complete, relevant description information of p...	0.815851
Complete, relevant description information of p...	Easy to use website or application	0.815851
Which device do you use to access the online sh...	What is the operating system (OS) of your device?	0.815200
What is the operating system (OS) of your device?	Which device do you use to access the online sh...	0.815200
Availability of several payment options	Presence of online assistance through multi-cha...	0.805550
Presence of online assistance through multi-cha...	Availability of several payment options	0.805550
Complete, relevant description information of p...	Availability of several payment options	0.765623
Availability of several payment options	Complete, relevant description information of p...	0.765623
Longer delivery period	Longer page loading time (promotion, sales period)	0.763026
Longer page loading time (promotion, sales period)	Longer delivery period	0.763026
Trust that the online retail store will fulfill...	You feel gratification shopping on your favorit...	0.762427
You feel gratification shopping on your favorit...	Trust that the online retail store will fulfill...	0.762427
Easy to use website or application	Presence of online assistance through multi-cha...	0.755796
Presence of online assistance through multi-cha...	Easy to use website or application	0.755796
	From the following, tick any (or all) of the on...	0.754655
From the following, tick any (or all) of the on...	Presence of online assistance through multi-cha...	0.754655
Complete, relevant description information of p...	Ease of navigation in website	0.753915
Ease of navigation in website	Complete, relevant description information of p...	0.753915
The content on the website must be easy to read...	User derive satisfaction while shopping on a go...	0.751771
User derive satisfaction while shopping on a go...	The content on the website must be easy to read...	0.751771

Correlation with Heatmap:

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.Correlation statistics can be used in finance and investing. Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.

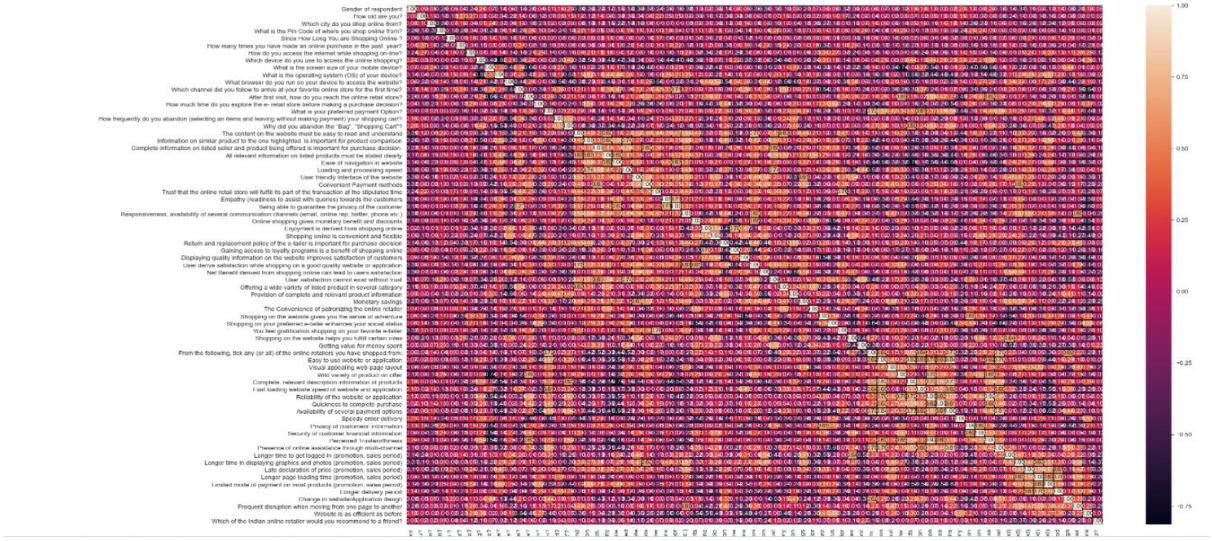
It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the **matrix data structure** is used. Such matrix is called as **correlation matrix.**

Correlation heatmap is graphical representation of **correlation matrix** representing correlation between different variables.

For to do feature selection and make feature ready for the model building.we check correlation of variables using heatmap.And describe method for the census data set.

```
In [68]: plt.figure(figsize=(25,15))
         sns.heatmap(df1.corr(),annot=True,linewidths=0.5,linecolor="black",fmt='.2f')
```

Out[68]: <AxesSubplot:>



the highly positively correlated column is ' Getting value for money spent' and the most negatively correlated column is 'Longer time in displaying graphics and photos (promotion, sales period)' .the 'Perceived Trustworthiness' makes zero correlation with the target column

```
In [69]: df1.describe()
```

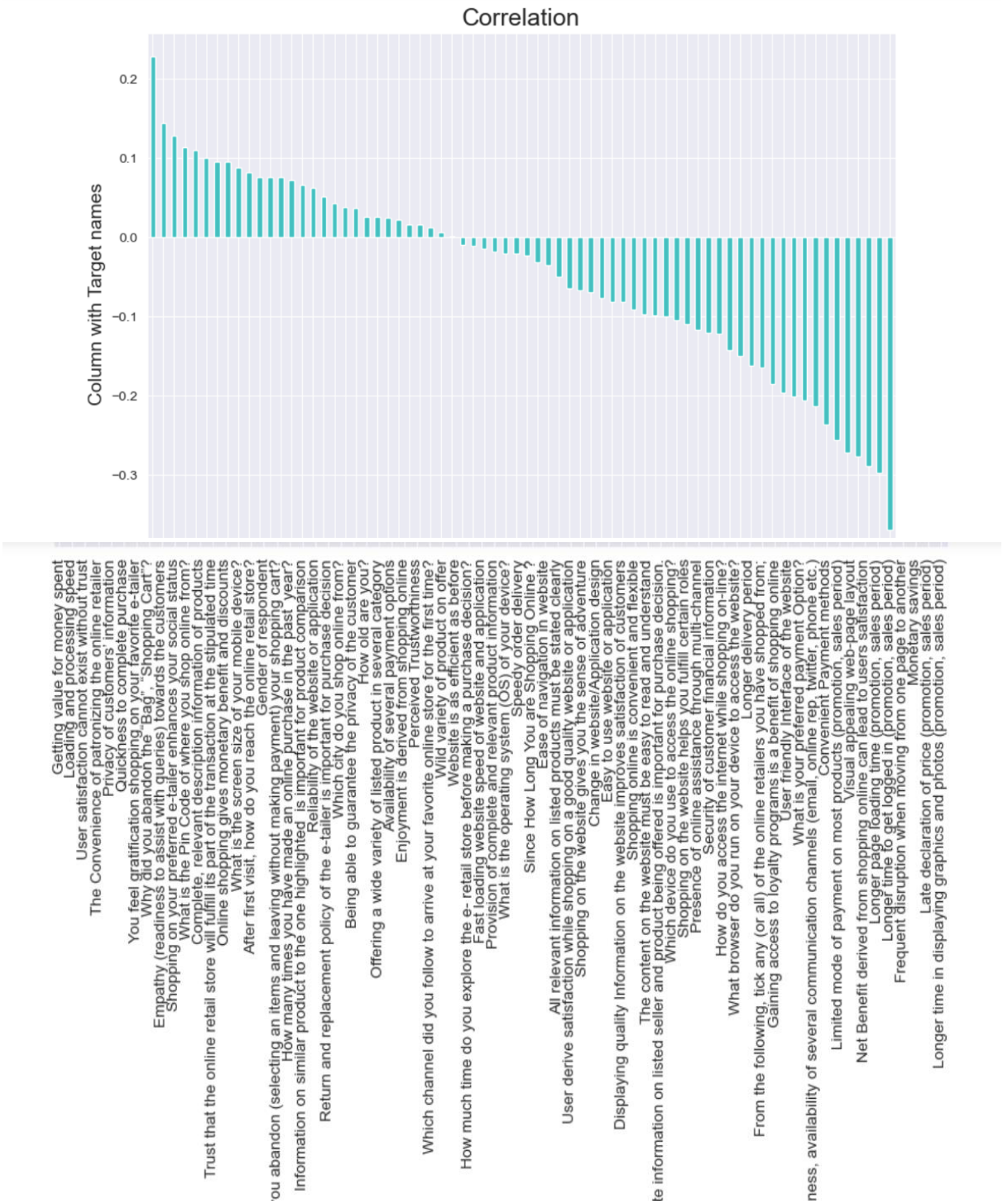
Out[69]:

	Gender of respondent	How old are you?	Which city do you shop online from?	What is the Pin Code of where you shop online from?	Since How Long You are Shopping Online ?	How many times you have made an online purchase in the past year?	How do you access the internet while shopping on-line?	Which device do you use to access the online shopping?	What is the screen size of your mobile device?	What is the operating system (OS) of your device?	What browser do you run on your device to access the website?	Which channel did you follow to arrive at your favorite online store for the first time?
count	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000	477.000000
mean	0.331237	1.320755	4.404612	211349.985325	2.371069	3.071279	2.155136	1.639413	2.228512	1.048218	0.547170	1.744235
std	0.471152	1.189877	2.977122	132053.474386	1.149827	1.778476	0.689842	0.730458	0.968075	0.859806	1.135919	0.581213
min	0.000000	0.000000	0.000000	110008.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	2.000000	122018.000000	1.000000	2.000000	2.000000	1.000000	2.000000	0.000000	0.000000	2.000000
50%	0.000000	1.000000	4.000000	201305.000000	3.000000	3.000000	2.000000	2.000000	2.000000	1.000000	0.000000	2.000000
75%	1.000000	2.000000	6.000000	201310.000000	3.000000	5.000000	3.000000	2.000000	3.000000	2.000000	0.000000	2.000000
max	1.000000	4.000000	10.000000	560037.000000	4.000000	5.000000	3.000000	3.000000	3.000000	2.000000	3.000000	2.000000

since these are all categorical columns there is no much use of outliers and skewness here

Correlation model:

Graph depicts clearly the positive and negative correlation of each variables with target column, justifies the outcome outlined in Multivariate analysis, that higher the education higher the gain & vice-versa



These are positively correlated columns

Getting value for money spent Loading and processing speed User satisfaction cannot exist without trust The Convenience of patronizing the online retailer Privacy of customers' information Quickness to complete purchase You feel gratification shopping on your favorite e-tailer

Below are negatively correlated columns

Longer time in displaying graphics and photos (promotion, sales period)', Late declaration of price (promotion, sales period) Monetary savings Frequent disruption when moving from one page to another

this graph shows the positive and negative correlation of each variables with target column

In [71]:
```
#VIF calculation
import statsmodels.api as sm
from scipy import stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

In [72]:
```
df1.shape
```

Out[72]: (477, 71)

In [73]:
```
df1.shape[1]
```

Out[73]: 71

In [74]:
```
#calculates vif
def calc_vif(df1):
    vif=pd.DataFrame()
    vif['Variables']=df1.columns
    vif['VIF FACTOR']=[variance_inflation_factor(df1.values,i)for i in range(df1.shape[1])]
    return(vif)


calc_vif(df1)
```

Out[74]:

	Variables	VIF FACTOR
9	What is the operating system (OS) of your device?	34.601079
10	What browser do you run on your device to acce...	inf
11	Which channel did you follow to arrive at your...	inf
12	After first visit, how do you reach the online...	inf
13	How much time do you explore the e- retail sto...	inf
14	What is your preferred payment Option?	inf
15	How frequently do you abandon (selecting an it...	inf
16	Why did you abandon the “Bag”, “Shopping Cart”?	inf
17	The content on the website must be easy to rea...	inf
18	Information on similar product to the one high...	inf
19	Complete information on listed seller and prod...	inf
20	All relevant information on listed products mu...	inf
21	Ease of navigation in website	inf

Some of the vif values are very high and since its infinity we use PCA principle Component Analysis Technique to make column reduction.

```
In [75]: df1=df1.drop(['Perceived Trustworthiness'],axis=1)
```

```
In [76]: calc_vif(df1)
```

Out[76]:

	Variables	VIF FACTOR
0	Gender of respondent	2.345747
1	How old are you?	1.911979
2	Which city do you shop online from?	1.659483
3	What is the Pin Code of where you shop online ...	2.012503
4	Since How Long You are Shopping Online ?	1.545749
5	How many times you have made an online purchas...	2.435755
6	How do you access the internet while shopping ...	2.310475
7	Which device do you use to access the online s...	40.336752
8	What is the screen size of your mobile device?	29.122274
9	What is the operating system (OS) of your device?	34.601079
10	What browser do you run on your device to acce...	inf

Dropped this column since it has zero correlation with the target.

Hardware and Software Requirements and Tools Used

HARDWARE & SOFTWARE TOOLS, LIBRARIES AND PACKAGES USED:

Hardware :Intel i7,RAM 16GB used.

Software: Jupyter Notebook (Anaconda 3)

Language: Python

Libraries:

1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Sklean
6. Scipy
7. Statsmodels
8. Pip-Package install Manager

```
#import all libaries
import pandas as pd
import numpy as np
import statistics
import seaborn as sns
from matplotlib import pyplot as plt
import sklearn
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
import warnings
warnings.filterwarnings('ignore')
url="https://raw.githubusercontent.com/dsrscientist/dataset1/master/census_income.csv"
```

Category	Tool	Function
Data loading and analysis	Import pandas as pd	Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. Pandas is built on top of numpy. So, numpy gets some superpower with pandas. It offers data structures and operations for manipulating numerical tables and time series.
	Import numpy as np	NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.It has Quantile method too for removing outliers. It is the fundamental package for scientific computing with Python
Data visualization	Import matplotlib.pyplot as plt	Matplotlib is a plotting library used for data visualization.
	Import seaborn as sns	Seaborn is also a plotting library. It is more advanced than matplotlib but works with matplotlib
Scikit Learn Preprocessing Libraries	Sklearn.preprocessing	Package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.Has power transformer to remove skewness. In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate. It has MinMaxScaler to scale the data.
	Sklearn.preprocessing import LabelEncoder	Label Encoding in Python can be implemented using the Sklearn Library. Sklearn furnishes a very effective method for encoding the categories of categorical features into numeric values. Label encoder encodes labels with credit between 0 and n-1 classes where n is the number of diverse labels.
Import statistics	Import statsmodels.api as sm	From scipy import stats This module provides functions for calculating mathematical statistics of numeric (Real-valued) data. This library provides a number of common functions and types useful in statistics. It focus on high performance, numerical robustness, and use of good algorithms

In [55]:
```
#VIF calculation
import statsmodels.api as sm
from scipy import stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library and we can import this library.

```
In [75]: #train test split
         from sklearn.model_selection import train_test_split

         x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=40)
```

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train_test_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train_test_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

The most commonly used Performance metrics for classification problem are as follows,

- Accuracy.
- Confusion Matrix.
- Precision, Recall, and F1 score.
- ROC AUC.
- Log-loss.

```
In [93]: from sklearn.linear_model import LogisticRegression
         from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
```

```
In [115]: # multi-class classification
          from sklearn.multiclass import OneVsRestClassifier
          from sklearn.metrics import roc_curve,auc
          from sklearn.metrics import roc_auc_score
```

```
#perform gridsearchcv and cross val score on LinearRegression
from sklearn.model_selection import GridSearchCV
```

Grid search is used as an approach to hyper-parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. GridSearchCV helps us combine an estimator with a grid search preamble to tune hyper-parameters.

```
from sklearn.model_selection import cross_val_score
```

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.
The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.

2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

```
In [87]: #importing and fitting the data to the pca
         from sklearn.decomposition import PCA
```

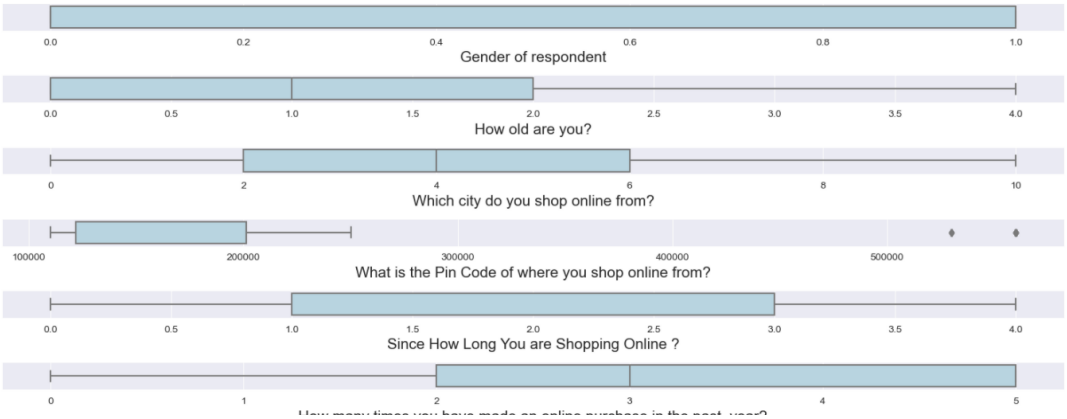
The most important use of PCA is to **represent a multivariate data table as smaller set of variables** (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables.

Also import all the required algorithms for classification purpose below.

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

```
In [77]: plt.figure(figsize= (15,20))
         pltnum = 1
         for i in df1:
             if pltnum <=20:
                 plt.subplot(20,1,pltnum)
                 sns.boxplot(df1[i],color = 'lightblue')
                 plt.xlabel(i,fontsize=15)
             pltnum+=1
         plt.tight_layout()
```



In [78]:
```
df1.skew()
```

Out[78]:
```
Gender of respondent                                                                                  0.719401
How old are you?                                                                                      0.599562
Which city do you shop online from?                                                                   0.377239
What is the Pin Code of where you shop online from?                                                   1.940416
Since How Long You are Shopping Online ?                                                             -0.302447
How many times you have made an online purchase in the past  year?                                   -0.342079
How do you access the internet while shopping on-line?                                               -0.367625
Which device do you use to access the online shopping?                                               -0.297123
What is the screen size of your mobile device?                                                       -1.280122
What is the operating system (OS) of your device?                                                    -0.092788
What browser do you run on your device to access the website?                                         1.644724
Which channel did you follow to arrive at your favorite online store for the first time?             -2.155916
After first visit, how do you reach the online retail store?                                         -0.033138
How much time do you explore the e- retail store before making a purchase decision?                  -0.447312
What is your preferred payment Option?                                                               -0.018479
How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart?  -0.855688
Why did you abandon the “Bag”, “Shopping Cart”?                                                       0.552623
The content on the website must be easy to read and understand                                       -0.725199
Information on similar product to the one highlighted  is important for product comparison           -0.116210
Complete information on listed seller and product being offered is important for purchase decision.   0.026120
All relevant information on listed products must be stated clearly                                    0.229556
Ease of navigation in website                                                                        -0.178049
Loading and processing speed                                                                          0.160018
User friendly Interface of the website                                                               -1.500022
Convenient Payment methods                                                                           -0.831567
Trust that the online retail store will fulfill its part of the transaction at the stipulated time   -0.236387
```

since these are all categorical columns and of discrete values.Hence no need to care much about outliers and skewness.Because it applies to numerical data.

#checking Z-score to remove outliers

In [79]:
```
import numpy as np
from scipy.stats import zscore
z=np.abs(zscore(df1))
z.shape
```

Out[79]:
```
(477, 70)
```

In [80]:
```
threshold=3
print(np.where(z>3))
```

```
(array([  0,   8,   9,  32,  33,  42,  43,  60,  61,  70,  71,  97,  98,
       144, 151, 152, 161, 164, 178, 179, 191, 197, 206, 207, 213, 214,
       226, 232, 241, 242, 258, 259, 291, 292, 295, 296, 320, 325, 328,
       353, 354, 357, 360, 365, 373, 376, 414, 415, 418, 419, 445, 469,
       470, 474], dtype=int64), array([ 6, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 26, 11, 11, 26,
       11, 11, 11, 26, 26, 11,  6, 11, 11, 26, 26, 11,  6, 11, 11, 11, 11,
       11, 11, 26, 26, 26, 26, 11, 26, 26, 11, 26, 26, 11, 11, 11, 11,  6,
       11, 11, 26], dtype=int64))
```

In [81]:
```
df_new=df1[(z<3).all(axis=1)]
print(df1.shape)
print(df_new.shape)
```

```
(477, 70)
(423, 70)
```

some outliers got removed

In [82]:
```
#finds data loss
loss_percent=(477-423)/(477*100)
print(loss_percent)
```

```
0.0011320754716981133
```

since data percentage loss is too low and its cleaned now

In [83]:
```
df_new.shape
```

Out[83]:
```
(423, 70)
```

In [84]:
```
#seggregate input data and output data
x=df_new.iloc[:,:-1]
y=df_new.iloc[:,-1]
```

SKEWNESS REMOVAL AND SCALING

In [85]:
```
#removing skewness
from sklearn.preprocessing import power_transform
x=power_transform(x,method='yeo-johnson')
x
```

```
[-0.71842121, -1.29961483, -0.07067973, ...,  1.05567371,
  0.84949321, -1.26406479],
...,
[-0.71842121, -0.11530563, -0.44954415, ..., -0.2115564 ,
 -0.88953007,  0.96568382],
[-0.71842121,  1.77544937,  1.69458292, ..., -1.41183213,
  0.40675544,  1.20382923],
[-0.71842121,  0.67171772, -0.44954415, ..., -1.41183213,
 -1.70925628, -1.26406479]])
```

SCALING

In [86]:
```
#scaling to get better model performance
from sklearn.preprocessing import MinMaxScaler
mmscaler = MinMaxScaler()
x = mmscaler.fit_transform(x)
x
```

Out[86]:
```
array([[0.        , 0.        , 0.30723766, ..., 0.        , 0.55828533,
        0.34851377],
       [0.        , 0.        , 0.30723766, ..., 0.        , 0.55828533,
        0.34851377],
       [0.        , 0.        , 0.52425575, ..., 1.        , 0.85135035,
        0.        ],
       ...,
       [0.        , 0.38513316, 0.42215051, ..., 0.48643278, 0.27274034,
        0.7856909 ],
       [0.        , 1.        , 1.        , ..., 0.        , 0.70404208,
        0.8696056 ],
       [0.        , 0.64107037, 0.42215051, ..., 0.        , 0.        ,
        0.        ]])
```

PCA

PRINCIPLE COMPONENT ANALYSIS

In [87]:
```
#importing and fitting the data to the pca
from sklearn.decomposition import PCA
```

In [88]:
```
pca=PCA(n_components=65)
pca.fit(x)
```

Out[88]:
```
PCA(n_components=65)
```

In [89]:
```
#checking the explained variance by the principal component
pca.explained_variance_ratio_
```

Out[89]:
```
array([1.87161333e-01, 1.52111294e-01, 1.22219935e-01, 9.88658418e-02,
       8.57384067e-02, 5.25452695e-02, 4.24292453e-02, 3.55558697e-02,
       3.52004029e-02, 2.99435247e-02, 2.90157026e-02, 2.64526674e-02,
       2.25355975e-02, 1.78966395e-02, 1.54271226e-02, 1.02597339e-02,
       8.96299419e-03, 6.86933807e-03, 6.24266310e-03, 5.42686847e-03,
       3.87454387e-03, 3.21895554e-03, 1.52665894e-03, 3.52876719e-04,
       1.66514813e-04, 3.69110647e-32, 1.64759172e-32, 1.16416983e-32,
       1.04845475e-32, 9.84052719e-33, 7.72977483e-33, 7.19402720e-33,
       6.77924627e-33, 6.09032120e-33, 5.36772445e-33, 4.16458482e-33,
       4.06838299e-33, 3.55933644e-33, 2.54097204e-33, 1.96761419e-33,
       1.74380056e-33, 1.10233997e-33, 1.10233997e-33, 1.10233997e-33,
       1.10233997e-33, 1.10233997e-33, 1.10233997e-33, 1.10233997e-33,
       1.10233997e-33, 1.10233997e-33, 1.10233997e-33, 1.10233997e-33,
       1.10233997e-33, 1.10233997e-33, 1.10233997e-33, 1.10233997e-33,
       1.10233997e-33, 1.10233997e-33, 1.10233997e-33, 1.10233997e-33,
       1.10233997e-33, 1.10233997e-33, 1.10233997e-33, 1.10233997e-33,
```

In [90]:
```
#transforming the principle component to the original values
x_returned_pca=pca.transform(x)
```

In [91]:
```
x_returned_pca.shape
```

Out[91]:
```
(423, 65)
```

Testing of Identified Approaches (Algorithms)

These are all the Algorithms used for Model Building and Prediction.We did Hyper Parameter Tuning with these algorithms using the GridSearchCV.

RandomForestClassifier
GaussianNB
KNeighborsClassifier
AdaBoost Classifier
SVC(Support Vector Classifier)
LogisticRegression
Soft Voting Classifier
Hard Voting classifier
DecisionTreeClassifier
Gradient Boosting Classifier
LightGradientBoostingClassifier
CatBoostClassifier

ExtraTreesClassifier
XGBoost Classifier

These algorithms has been used for both Training and Testing purpose and got evaluated with classification metrics such as f1score,confusion matrix,precision,recall and AUC ROC curve etc.,

Run and Evaluate selected models

Logistic Regression:

- Logistic regression is used **to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables**. Logistic regression is a **simple and more efficient method for binary and linear classification problems**. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification.Logistic Regression is used **when the dependent variable(target) is categorical**.

MODEL PREDICTION

LOGSITIC REGRESSION

In [92]:
```
#train test split
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=40)
```

In [93]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
from sklearn.model_selection import train_test_split
```

Hyperparameter Tuning Done Using GridSearchCV:

parameter tuning

In [97]:
```
#performs GridsearchCV logistic regression
from sklearn.model_selection import GridSearchCV
parameters={'dual':[False,True],'fit_intercept':[True,False],'random_state':list(range(0,1)),'max_iter':[100,50], 'tol':[0.001,0.
lr=LogisticRegression()
clf=GridSearchCV(lr,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'dual': False, 'fit_intercept': False, 'max_iter': 50, 'random_state': 0, 'tol': 0.001}
```

In [98]:
```
lr=LogisticRegression(fit_intercept= False, dual=False, max_iter= 50, random_state=0, tol= 0.001)
lr.fit(x_train,y_train)
pred_test_lr=lr.predict(x_test)
pred_train_lr=lr.predict(x_train)
lr_score = lr.score(x_train,y_train)
lr_acc_score=accuracy_score(y_test,pred_test)
print("Accuracy score is:",lr_acc_score*100)
print("score of model is:",lr_score*100)
```

```
Accuracy score is: 42.35294117647059
score of model is: 57.100591715976336
```

In [99]:
```
cv_score_lr=cross_val_score(lr,x,y,cv=5)
cv_mean_lr=cv_score_lr.mean()
print("cv_mean is:",cv_mean_lr*100)
```

```
cv_mean is: 50.36134453781512
```

In [100]:
```
print(classification_report(y_test, pred_test))
```

```
              precision    recall  f1-score   support

           0       0.14      0.07      0.10        14
           1       0.14      0.33      0.20         9
           2       0.00      0.00      0.00         9
           3       0.00      0.00      0.00         2
           4       0.56      0.60      0.58        40
           5       1.00      0.73      0.84        11

    accuracy                           0.42        85
   macro avg       0.31      0.29      0.29        85
weighted avg       0.43      0.42      0.42        85
```

In [101]:
```
print(confusion_matrix(y_test,pred_test))
```

```
[[ 1  3  0  0 10  0]
 [ 1  3  2  0  3  0]
 [ 0  3  0  0  6  0]
 [ 0  1  1  0  0  0]
 [ 5 11  0  0 24  0]
 [ 0  0  3  0  0  8]]
```

In [102]:
```
print(accuracy_score(y_test,pred_test)*100)
```

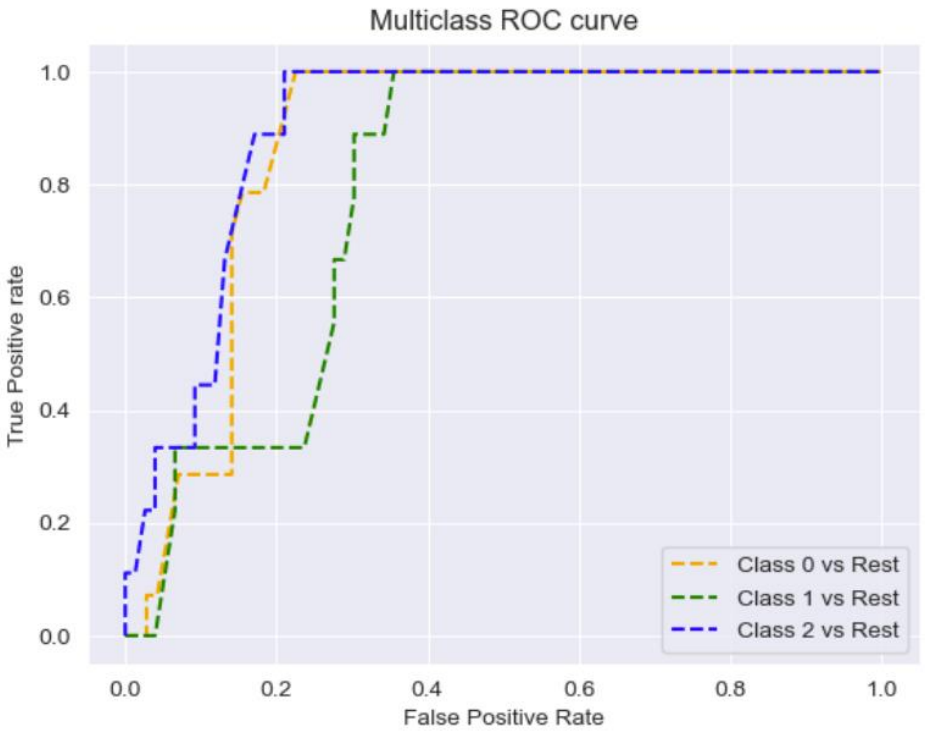
```
42.35294117647059
```

AUc-ROC Curve

In [103]:
```
!pip install -U scikit-learn
```

```
Requirement already satisfied: scikit-learn in c:\users\srividya\anaconda3\lib\site-packages (1.0.1)
Requirement already satisfied: numpy>=1.14.6 in c:\users\srividya\anaconda3\lib\site-packages (from scikit-learn) (1.20.1)
Requirement already satisfied: joblib>=0.11 in c:\users\srividya\anaconda3\lib\site-packages (from scikit-learn) (1.0.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\srividya\anaconda3\lib\site-packages (from scikit-learn) (2.1.
0)
Requirement already satisfied: scipy>=1.1.0 in c:\users\srividya\anaconda3\lib\site-packages (from scikit-learn) (1.6.2)
```

In [104]:
```
# multi-class classification
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import roc_curve,auc
from sklearn.metrics import roc_auc_score

# generate 2 class dataset
#x, y = make_classification(n_samples=1000, n_classes=3, n_features=20, n_informative=3, random_state=42)

# fit model
clf = OneVsRestClassifier(LogisticRegression())
clf.fit(x_train, y_train)
pred = clf.predict(x_test)
pred_prob = clf.predict_proba(x_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(y_test, pred_prob[:,i], pos_label=i)

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class 0 vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class 1 vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue',label='Class 2 vs Rest')

plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);
```



RANDOM FOREST CLASSIFIER

The random forest is a **classification algorithm consisting of many decisions trees**. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. It **can perform both regression and classification tasks**. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.
Random forest **adds additional randomness to the model, while growing the trees**. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Parameter Tuning:

In [110]:
```
#performs GridSearchCV on RandomForestClassifier
from sklearn.model_selection import GridSearchCV
parameters={'criterion':['gini', 'entropy'],'n_estimators':[50,100],'max_features':['auto', 'sqrt', 'log2'],'random_state':list(r
rf=RandomForestClassifier()
clf=GridSearchCV(rf,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'bootstrap': True, 'criterion': 'gini', 'max_features': 'log2', 'min_weight_fraction_leaf': 0.1, 'n_estimators': 50, 'random_s
tate': 0}
```

In [111]:
```
estClassifier(criterion="gini",max_features="log2",n_estimators=50,random_state=0,bootstrap="True",min_weight_fraction_leaf= 0.1
ain,y_train)
f=rf.predict(x_test)
f=rf.predict(x_train)
f.score(x_train,y_train)
=accuracy_score(y_test,pred_test)
acy score is:",rf_acc_score*100)
 of model is:",rf_score*100)
```

```
Accuracy score is: 42.35294117647059
score of model is: 52.071005917159766
```

In [112]:
```
cv_score_rf=cross_val_score(rf,x,y,cv=5)
cv_mean_rf=cv_score_rf.mean()
print("cv_mean is:",cv_mean_rf*100)
```

```
cv_mean is: 51.53501400560223
```

In [113]:
```
print(confusion_matrix(y_test,pred_test_rf))
```

```
[[ 0  3  0  0 11  0]
 [ 0  1  0  0  6  2]
 [ 0  0  0  0  9  0]
 [ 0  0  0  0  0  2]
 [ 0  4  0  0 36  0]
 [ 0  0  0  0  2  9]]
```

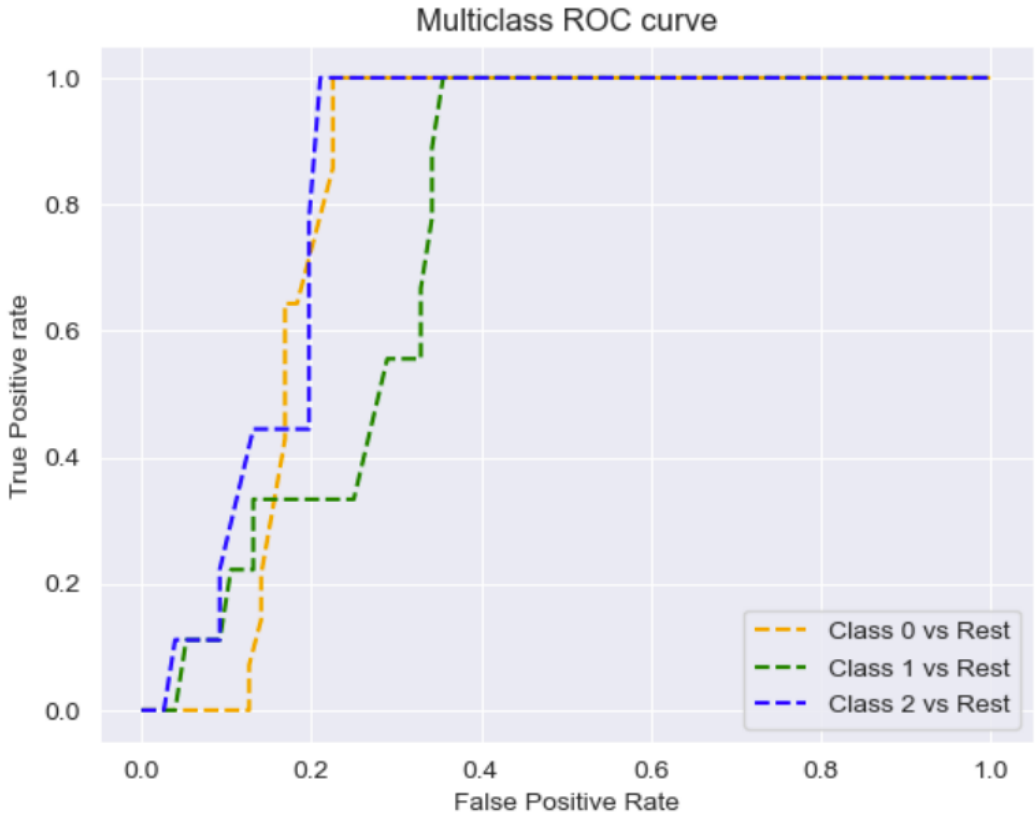
In [114]:
```
print(classification_report(y_test,pred_test_rf))
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	14
1	0.12	0.11	0.12	9
2	0.00	0.00	0.00	9
3	0.00	0.00	0.00	2
4	0.56	0.90	0.69	40
5	0.69	0.82	0.75	11
accuracy			0.54	85
macro avg	0.23	0.30	0.26	85
weighted avg	0.37	0.54	0.44	85

AUC-ROC CURVE:

In [115]:
```
# multi-class classification
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import roc_curve,auc
from sklearn.metrics import roc_auc_score

# generate 2 class dataset
#x, y = make_classification(n_samples=1000, n_classes=3, n_features=20, n_informative=3, random_state=42)

# fit model
clf = OneVsRestClassifier(RandomForestClassifier())
clf.fit(x_train, y_train)
pred = clf.predict(x_test)
pred_prob = clf.predict_proba(x_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(y_test, pred_prob[:,i], pos_label=i)

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class 0 vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class 1 vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue',label='Class 2 vs Rest')

plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
```

```
plt.savefig('Multiclass ROC',dpi=300);
```



DECISION TREE CLASSIFIER

The main advantage of the decision tree classifier is **its ability to using different feature subsets and decision rules at different stages of classification**. Decision tree often involves higher time to train the model. Decision tree training is relatively expensive as the complexity and time has taken are more. The Decision Tree algorithm **is inadequate for applying regression and predicting continuous values**.In this, **the data is continuously split according to a certain parameter**. The tree can be explained by two entities, namely decision nodes and leaves.

parameter tuning

In [118]:
```
#perform gridsearchcv and cross val score on Decison Tree DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
parameters={'criterion':['gini', 'entropy'],'splitter':['best','random'],'max_features':['auto', 'sqrt', 'log2'],'random_state':1
dt=DecisionTreeClassifier()
clf=GridSearchCV(dt,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'criterion': 'gini', 'max_depth': 11, 'max_features': 'log2', 'min_samples_leaf': 3, 'min_samples_split': 2, 'random_state':
5, 'splitter': 'random'}
```

In [119]:
```
iterion='gini',max_features= 'log2',max_depth= 11 , random_state= 5, splitter= 'random',min_samples_leaf= 3,min_samples_split= 2)

st)
rain)
/_train)
/_test,pred_test_dt)
t_acc_score*100)
t_score*100)
```

```
Accuracy score is: 42.35294117647059
score of model is: 57.396449704142015
```

In [120]:
```
cv_score_dt=cross_val_score(dt,x,y,cv=5)
cv_mean_dt=cv_score_dt.mean()
print("cv_mean is:",cv_mean_dt*100)
```

```
cv_mean is: 49.89075630252101
```

In [121]:
```
print(confusion_matrix(y_test,pred_test_dtc))
```

```
[[ 1  3  0  0 10  0]
 [ 1  2  3  0  3  0]
 [ 0  4  1  0  4  0]
 [ 0  1  1  0  0  0]
 [10 11  2  0 17  0]
 [ 0  2  1  0  0  8]]
```

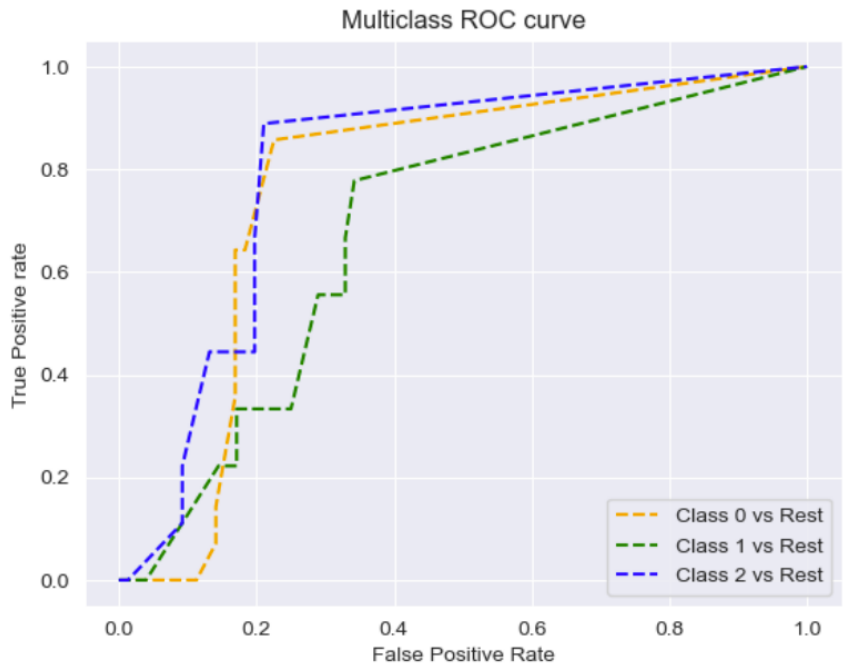
In [122]:
```
print(classification_report(y_test,pred_test_dtc))
```

```
              precision    recall  f1-score   support

           0       0.08      0.07      0.08        14
           1       0.09      0.22      0.12         9
           2       0.12      0.11      0.12         9
           3       0.00      0.00      0.00         2
           4       0.50      0.42      0.46        40
           5       1.00      0.73      0.84        11

    accuracy                           0.34        85
   macro avg       0.30      0.26      0.27        85
weighted avg       0.40      0.34      0.36        85
```

AUC-ROC CURVE:

In [123]:
```
# multi-class classification
from sklearn.multiclass import OneVsRestClassifier
from sklearn.metrics import roc_curve,auc
from sklearn.metrics import roc_auc_score

# generate 2 class dataset
#x, y = make_classification(n_samples=1000, n_classes=3, n_features=20, n_informative=3, r

# fit model
clf = OneVsRestClassifier(DecisionTreeClassifier())
clf.fit(x_train, y_train)
pred = clf.predict(x_test)
pred_prob = clf.predict_proba(x_test)

# roc curve for classes
fpr = {}
tpr = {}
thresh ={}

n_class = 3

for i in range(n_class):
    fpr[i], tpr[i], thresh[i] = roc_curve(y_test, pred_prob[:,i], pos_label=i)

# plotting
plt.plot(fpr[0], tpr[0], linestyle='--',color='orange', label='Class 0 vs Rest')
plt.plot(fpr[1], tpr[1], linestyle='--',color='green', label='Class 1 vs Rest')
plt.plot(fpr[2], tpr[2], linestyle='--',color='blue',label='Class 2 vs Rest')
```

```
plt.title('Multiclass ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.savefig('Multiclass ROC',dpi=300);
```



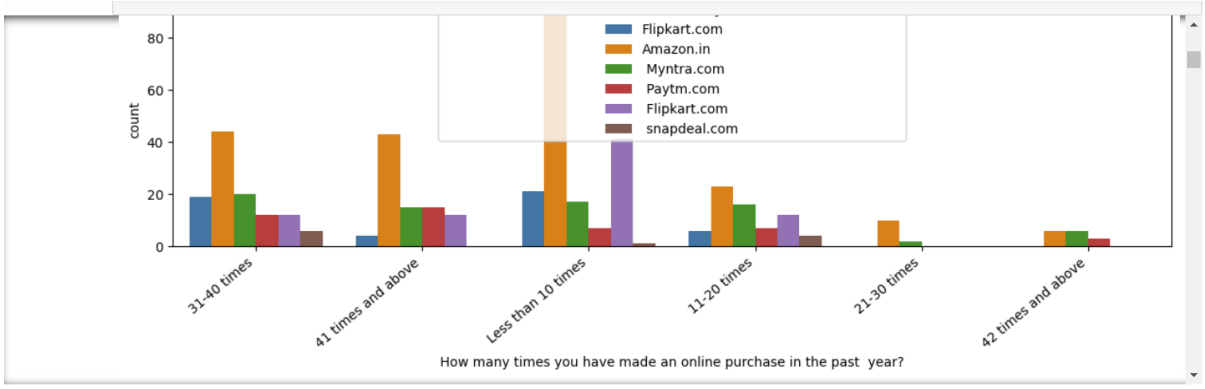
These are some of the algorithms used and it described here with the snapshot of their code and the results observed over different evaluation metrics are also mentioned.
The evaluation metrics used here is classification metrics.

Key Metrics for success in solving problem under consideration

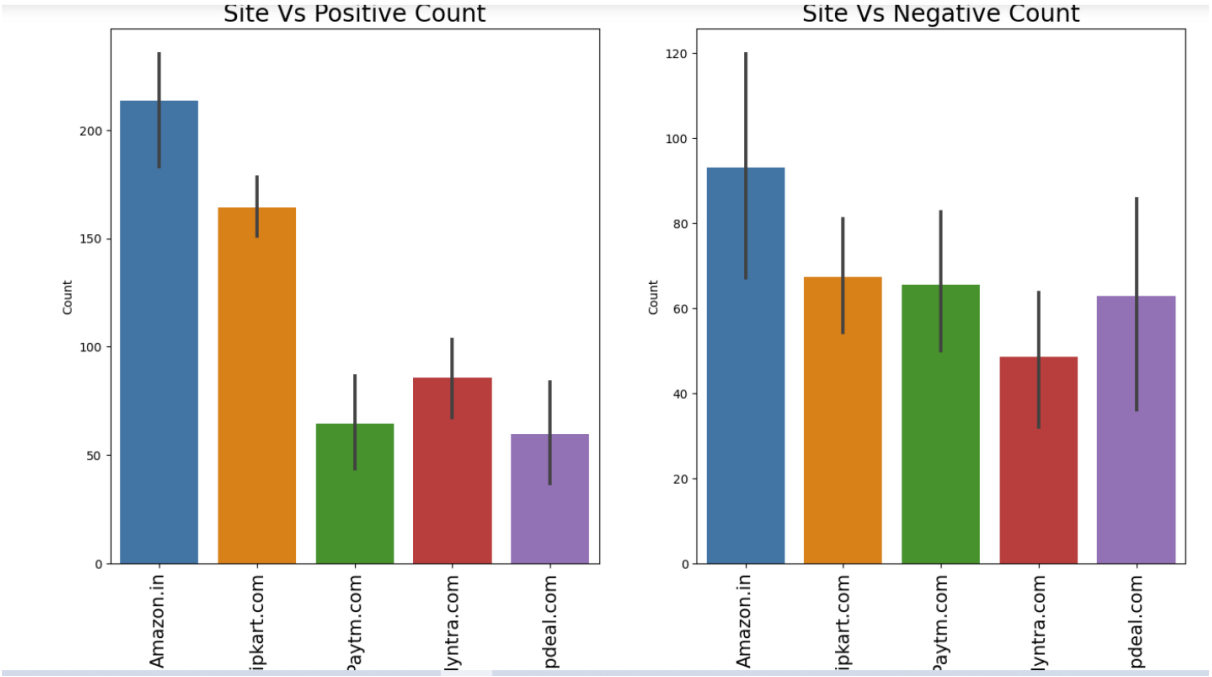
An evaluation metric **quantifies the performance of a predictive model**. This typically involves training a model on a dataset, using the model to make predictions on a holdout dataset not used during training, then comparing the predictions to the expected values in the holdout dataset. We got Good accuracy with Random Forest Classifier when comparing with other model's performance.

1. Random Forest reduces overfitting in decision trees and helps to improve the accuracy.So it gives good accuracy with our evaluation metric when used.
2. It is flexible to both classification and regression problems
3. It works well with both categorical and continuous values
4. It automates missing values present in the data
5. Normalising of data is not required as it uses a rule-based approach.

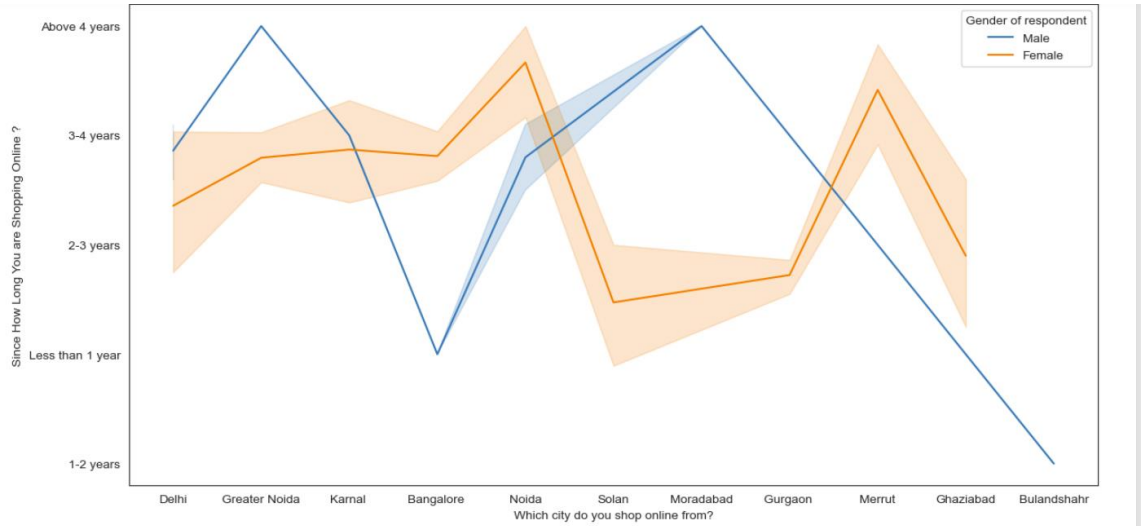
Visualizations



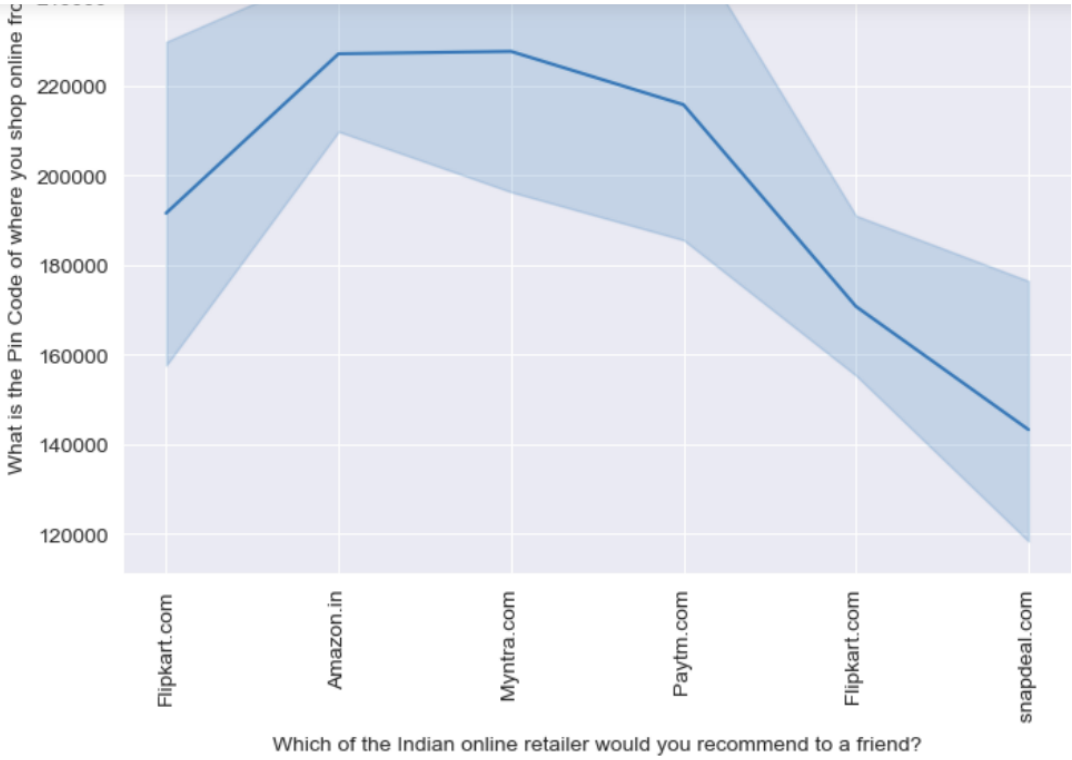
1.female shops most in amazon 2.the most shopping age is 21-30 years in amazon 3.most people are shopping from greater noida in amazon 4.people in 201305 pincode purchases more in amazon 5.people shopping around 4 years in amazon only 6.people made purchase in past years in amazon most 7.people use mobile internet and smartphone most to buy in amazon 8.screensize of mobile is 5.5 and has windows OS mostly 9.people use google chrome and search engine channel to explore first mostly 10.people purchase via application and direct url mostly 11.customer use credit/debit cards mostly for transaction in amazon 12.they abandon purchase sometimes only 13.they quit purchase because of better alternative offer 14.people strongly agree that the content of the website should be readable easily. 15.people mostly strongly agree to the both hedonic and utilitarian values provided by the e-retail 16.amazon and flipkart provides both hedonic and utilitarian values to the customers to make activation and retention.And they do best among others here



1.female shops more than males overall 2.the most shopping age is 31-40 years 3.most people are shopping from delhi in overall 4.people shopping around 4 years 5.people made purchase in past years in less than 10 times 6.people use mobile internet and smartphone most to buy in amazon 7.screensize of mobile is 5.5and others category and has windows OS mostly 8.people use google chrome and search engine channel to explore first mostly 9.people purchase via application and direct url mostly 10.people explore more than 15 minutes before their purchase 11.customer use credit/debit cards mostly for transaction in amazon 12.they abandon purchase sometimes only 13.they quit purchase because of better alternative offer and promo code if not available 14.people strongly agree that the content of the website should be readable easily. 15.people mostly strongly agree to the both hedonic and utilitarian values provided by the e-retail 16.amazon and flipkart provides both hedonic and utilitarian values to the customers to make activation and retention.And they do best among others here 17.myntra and paytm has longer page loading time 18.snapdeal and paytm has longer delivery speed 19.limitd mode of payments in snapdeal 20.amazon changed its webpage to upgraded one. 21.frequent disruption in snapdeal and paytm when navigation in page 22.amazon is efficient in webdesign as before



female purchases more than male



****

Amazon is the most recommanded webiste among all

Interpretation of the Results

CONCLUSION

OUTPUT

In [185]:
```
import numpy as np
a=np.array(y_test)
predicted=np.array(rf.predict(x_test))
df_con=pd.DataFrame({"Original":a,"Predicted":predicted},index=range(len(a)))
df_con
```

Out[185]:

	Original	Predicted
0	4	4
1	5	5
2	0	1
3	4	4
4	5	4
5	4	4
6	1	1
7	0	4
8	2	4
9	5	5
10	4	4

In [190]:
```
import pandas as pd
Model_scores=pd.DataFrame({})
Model_scores['Nos']=Nos
Model_scores['Model Names']=models
Model_scores['Scores']=scores
Model_scores.sort_values(by='Scores', ascending=False).style.hide_index()
```

Out[190]:

Nos	Model Names	Scores
2	RandomForestClassifier	54.117647
5	GaussianNB	54.117647
4	KNeighborsClassifier	48.235294
11	AdaBoost	47.058824
6	SVC	43.529412
1	LogisticRegression	42.352941
14	Soft Voting Classifier	38.823529
13	Voting classifier	36.470588
3	DecisionTreeClassifier	34.117647
7	Gradient Boosting Classifier	34.117647
8	Light Gradient Boosting Classifier	34.117647
9	CatBoostClassifier	34.117647
10	ExtraTreesClassifier	34.117647
12	XGBoost	34.117647

MODEL SAVING:

SAVE MODEL

In [186]:
```
import pickle
filename='E-retail_Customer_Activation_Retention_rf.pkl'
pickle.dump(rf,open(filename,'wb'))
```

In [192]:
```
df_con.to_csv("E-retail_Customer_Activation_Retention_rf.csv",sep='\t')
```

Inferences:

1.we can see that the amazon.in and flipkart.com are the e-commerce sites that most customer prefers for shopping

2.amazon is the most customer friendly,activated and retention website among others

3.amazon is the most shopped website among others for years

4.amazon is the most shopping website by the customers and they use credit or debit card for the transaction among other payment methods

5.the customers abandon the purchase before checkout because there are more better alternative offer available.And also the case when the promo code not applicable most of the times

6.Amazon and Flipkart are doing well in customer retention with hedonic values among other websites

7.Amazon is doing well in customer retention with utilitatian values among other websites

8.The top most good customer retention websites are amazon and flipkart

9.All the utilitarian values are provided mostly by amazon only when compared with others

Some findings are
1.female shops more than males overall
2.the most shopping age is 31-40 years
3.most people are shopping from delhi in overall
4.people shopping around 4 years
5.people made purchase in past years in less than 10 times
6.people use mobile internet and smartphone most to buy in amazon
7.screensize of mobile is 5.5and others category and has windows OS mostly
8.people use google chrome and search engine channel to explore first mostly
9.people purchase via application and direct url mostly
10.people explore more than 15 minutes before their purchase
11.customer use credit/debit cards mostly for transaction in amazon
12.they abandon purchase sometimes only
13.they quit purchase because of better alternative offer and promo code if not available
14.people strongly agree that the content of the website should be readable easily.
15.people mostly strongly agree to the both hedonic and utilitarian values provided by the e-retail

16.amazon and flipkart provides both hedonic and utilitarian values to the customers to make activation and retention.And they do best among others here
17.myntra and paytm has longer page loading time
18.snapdeal and paytm has longer delivery speed
19.limitd mode of payments in snapdeal
20.amazon changed its webpage to upgraded one.
21.frequent disruption in snapdeal and paytm when navigation in page
22.amazon is efficient in webdesign as before
Amazon is the most recommanded webiste among all

E-retail Retention exists most for amazon because
1.female shops most in amazon
2.the most shopping age is 21-30 years in amazon
3.most people are shopping from greater noida in amazon
4.people in 201305 pincode purchases more in amazon
5.people shopping around 4 years in amazon only
6.people made purchase in past years in amazon most
7.people use mobile internet and smartphone most to buy in amazon
8.screensize of mobile is 5.5 and has windows OS mostly
9.people use google chrome and search engine channel to explore first mostly
10.people purchase via application and direct url mostly
11.customer use credit/debit cards mostly for transaction in amazon
12.they abandon purchase sometimes only

13.they quit purchase because of better alternative offer
14.people strongly agree that the content of the website should be readable easily.
15.people mostly strongly agree to the both hedonic and utilitarian values provided by the e-retail
16.amazon and flipkart provides both hedonic and utilitarian values to the customers to make activation and retention.And they do best among others here

CONCLUSION

Key Findings and Conclusions of the Study

Customer Satisfaction and Retention can be improved by concentrating on the above mentioned hedonic and utilitarian values. And it can be attained by concentrating on the following impactful factors such as

Getting value for money spent Loading and processing speed User satisfaction cannot exist without trust The Convenience of patronizing the online retailer Privacy of customers' information Quickness to complete purchase You feel gratification shopping on your favorite e-tailer Longer time in displaying graphics and photos (promotion, sales period)', Late declaration of price (promotion, sales period) Monetary savings Frequent disruption when moving from one page to another
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

Learning Outcomes of the Study in respect of Data Science

From the above models Random Forest Classifier performs well. Because,The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. So we save this model for prediction

We faced multicollinearity issue because too many features in the dataset.Eventhough the number of features after reduced from 71 to

65.we still face multicollinearity issue.So we used PCA technique to overcome this And also RandomForest Classifier because it handles the problem of multicollinearity very well. Random Forest uses bootstrap sampling and feature sampling, i.e row sampling and column sampling.
Therefore **Random Forest is not affected by multicollinearity** that much since it is picking different set of features for different models and of course every model sees a different set of data points.
Correlated features will be given equal or similar importance, but overall reduced importance compared to the same tree built without correlated counterparts. Random Forests and decision trees, in general, **give preference to features with high cardinality** (Trees are biased to these type of variables.
And also It **reduces overfitting problem in decision trees** and also reduces the variance and therefore improves the accuracy.

Thus this Random Forest Classifier Model performs well for this dataset.so we saved this model.

Limitations of this work and Scope for Future Work

The main limitation of random forest is that **a large number of trees can make the algorithm too slow and ineffective for real-time predictions**. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained.

For **very large data sets, the size of the trees can take up a lot of memory**. It can tend to overfit, so you should tune the hyperparameters.

Here we used PCA to project the data into a new space where the 'new features' will be orthogonal to each other. We then, trained the model with the new features, but we found that the performance is the same.so we must You simply rotate original decision boundary to overcome this limitation in future and try with the same model.

For Implementation Code check my [srividya89/E-Retail-Customer-Retention (github.com)]