FLIP ROBO

E-retail Factors for Customer Activation and Retention

Submitted by:

Ganta Ganesh

ACKNOWLEDGMENT

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.

INTRODUCTION

When you buy a product or a service through internet, instead of going to a traditional brick-and mortar store, it is called online shopping. This trend from buying through internet is growing not only in India but, globally we see an increasing number of people are buying over the Internet because of its convenience. This year, holiday shopping for Christmas and New Year is estimated to be over $20 billion. In current scenario you can buy anything from net. You can purchase almost anything online starting with groceries, medicine, apparels, furniture, electronics, books, greeting cards to cell phones and ringtones for the cell phones, everything can be purchased online and what not. Still many people find it convenient to buy their groceries from the neighbourhood shop, many people are purchasing rail, air tickets and their holiday destinations over the Internet. Not only this but many people and corporate as well, are also purchasing a variety of services online such as a broking service, banking service and job search service.

Machine Learning

Machine learning is a capacity of the machines to analyse a set of data and build generic algorithms. There is no need to write codes, just feeding of data is enough it builds its logic based on it.

There are majorly two types of machine learning

- Supervised learning
- Un supervised learning

Supervised Learning

From the above context itself, it is so much evidence that there must be some supervisor as a teacher to carry on the process. Usually in this methodology, the training or the teaching is provided.

For Example, if it is to train to identify various kinds of fruits it must be like the shape of the fruit is round and a cavity is found at the top centre and the colour must be red. Which signifies Apple? If the shape is curved and long enough with the colour of green or yellow then it must be Banana. Now after this training it is given with another set of examples to identify.

Un Supervised Learning

This second type of learning in which there is no supervision or guidance required is called unsupervised learning. Here it can act without any guidance required. For example, if a picture of dogs and cats is given together to analyse, it has no information on Cats or Dogs. But still, it can categorize based on the similarities between them by analysing the patterns, Size, shape, figures, and differences.

Coverage of the Study

This report is restrained to the study of customer retention by online shopping platforms in India

Source of Data

The study is based on secondary data collected through various internet web sites.

Data Analysis

Analysis of data and the information collected from the secondary sources were made keeping the objectives of the study in mind.

When you buy a product or a service through internet, instead of going to a traditional brick-and mortar store, it is called online shopping. This trend from buying through internet is growing not only in India but, globally we see an increasing number of people are buying over the Internet because of its convenience. This year, holiday shopping for Christmas and New Year is estimated to be over $20 billion. In current scenario you can buy anything from net. You can purchase almost anything online starting with groceries, medicine, apparels, furniture, electronics, books, greeting cards to cell phones and ringtones for the cell phones, everything can be purchased online and what not. Still many people find it convenient to buy their groceries from the neighbourhood shop, many people are purchasing rail, air tickets and their holiday destinations over the Internet. Not only this but many people and corporate as well, are also purchasing a variety of services online such as a broking service, banking service and job search service.

Evolution of online shopping in India Online shopping had a rather slow and disorderly journey in India, it has not picked up as much as it should have primarily due to the fact that internet penetration itself was quite low and secondly (and importantly) the customers were not aware about it as well. Moreover, the customers are not ready to take the risk of buying a product without seeing it physically. Traditionally, Indians are conservative in their approach to shopping. They want to touch and feel the products and test its features before buying anything. Online shopping started early in 1995 by the introduction of internet in India. Online shopping became popular during the Internet boom in 1999-2000 with the well know auction site know as bazee.com. Soon amazon.com, the online bookstore founded by Jeff Bezos, created history by becoming the first bookstore with a presence only on the Internet. Later on, following the success of Amazon, many other bookstores with a physical presence also created an online presence on the Internet. Thereafter in 2005 bazee.com was taken up by eBay. The trend of online shopping took a good pace and many new portals started like amazon, flipkart, snapdeal, yebhi, gadgetsguru, myntra, iBibo, makemytrip, yatra, craftsvilla and so on. Many home portals such as Yahoo.com, Indiatimes.com and Rediff.com came up with online shopping options for the Indian consumer. It is convenient, faster and sometimes also cheaper than the traditional buying. Now a day‟s buying train ticket, bus ticket, air ticket all of them have gone through online option as well. Rather than standing in a long queue and waiting for your turn to purchase a ticket, people are finding it simpler to log on to a website and buy it. In some instances, you may have to pay a premium for an online purchase but it is still preferred because the convenience factor is much higher. For example, if you want to buy movie tickets online you may have to pay extra amount over the actual price of the ticket but because of its convenience, people are opting for it. Buying or placing an order online is also useful when you need to send a gift to a friend who is staying in a different city or country. For example, you can send flowers, cake and chocolates to your friend in New York on his/her birthday by placing an order for it on the Internet from your home in Mumbai.

Project Definition

To create a model that predicts the retention of customer by an online platform

Hardware and Software Requirements

SYSTEM SPECIFICATION

The hardware and the software specifications of the projects are
1) Hardware Requirements
Processor: Intel I 3
Ram: 4 GB
Hard disk Driver: 50 GB
Monitor: 15” Colour monitor
2) Software requirements
OS: Linux/ Windows/ MAC
Language: Python
Libraries: Jupyter notebook, Python, Matplot lib, Pandas, Numpy

PROJECT DESCRIPTION

Online shopping is a process by which the exchange of goods takes place over the internet where the buyer and seller come online to exchange goods. The customer finds a product on the website of the retailer and uses an online platform to buy the product. The main idea of online shopping is to reduce time.

Idea

To create a model that predicts the retention of customer by online platform
or negative.

Solution

Using classification models, we need to build a model that gives best predictions.

Summary

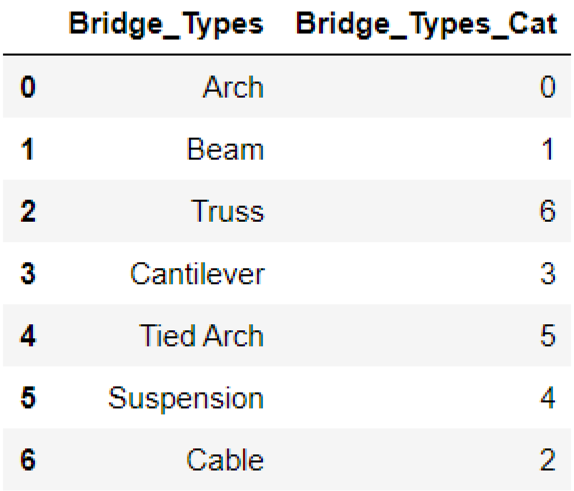
The ultimate goal of the project is to provide information on customer retention based on the features taken by various surveys

Classification

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.
Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy-to-understand example is classifying emails as “*spam*” or “*not spam*.”

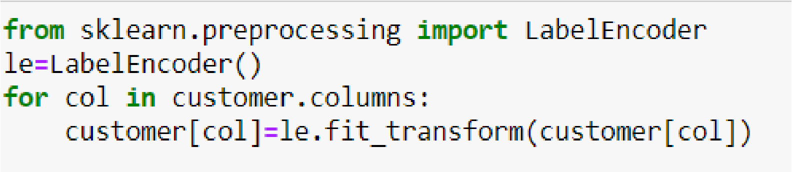
Label Encoding

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.



	Bridge_Types	Bridge_Types_Cat
0	Arch	0
1	Beam	1
2	Truss	6
3	Cantilever	3
4	Tied Arch	5
5	Suspension	4
6	Cable	2

Fig: Conversion of names to labels using Label encoding

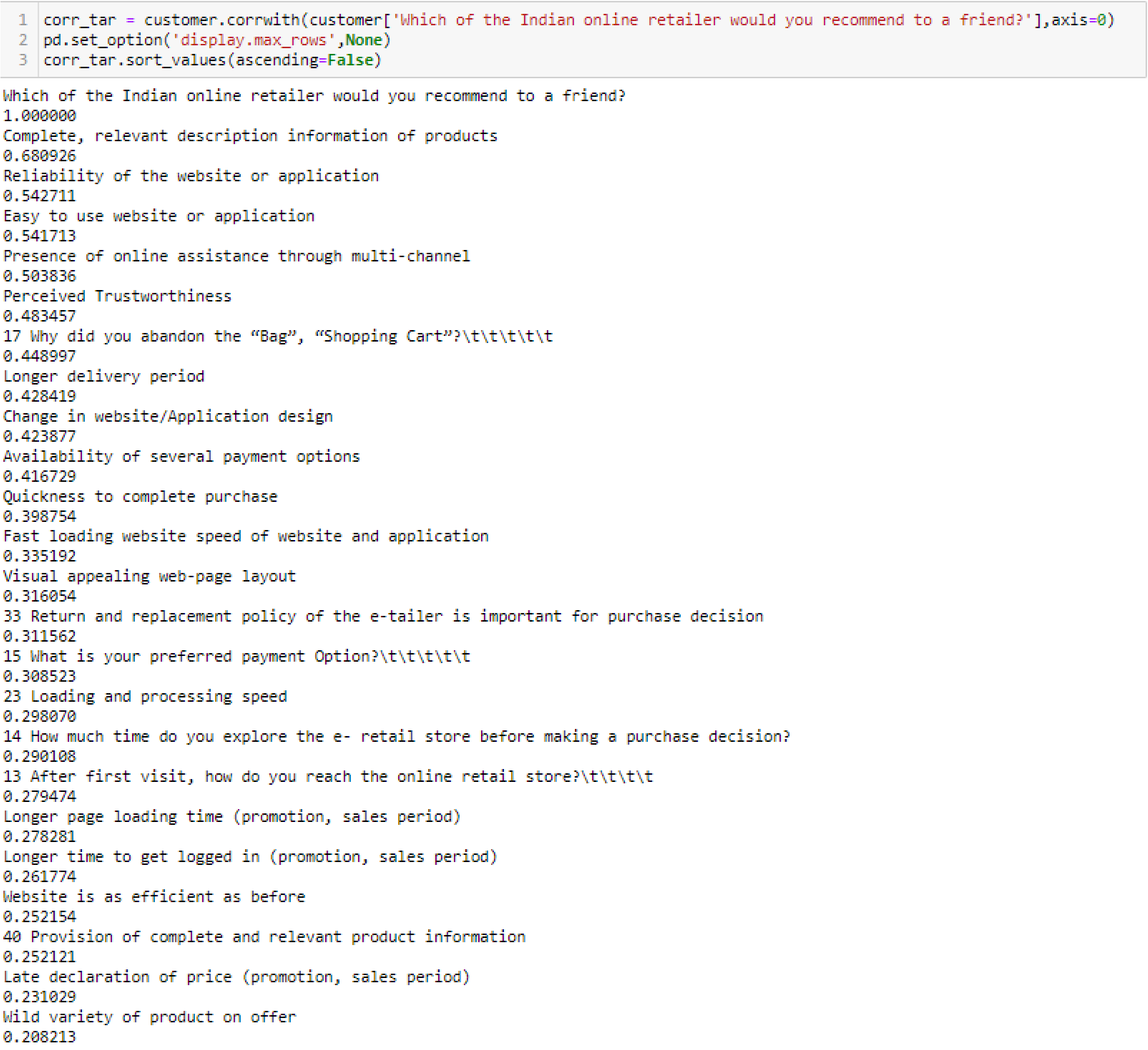


```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
for col in customer.columns:
    customer[col]=le.fit_transform(customer[col])
```

Correlation

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

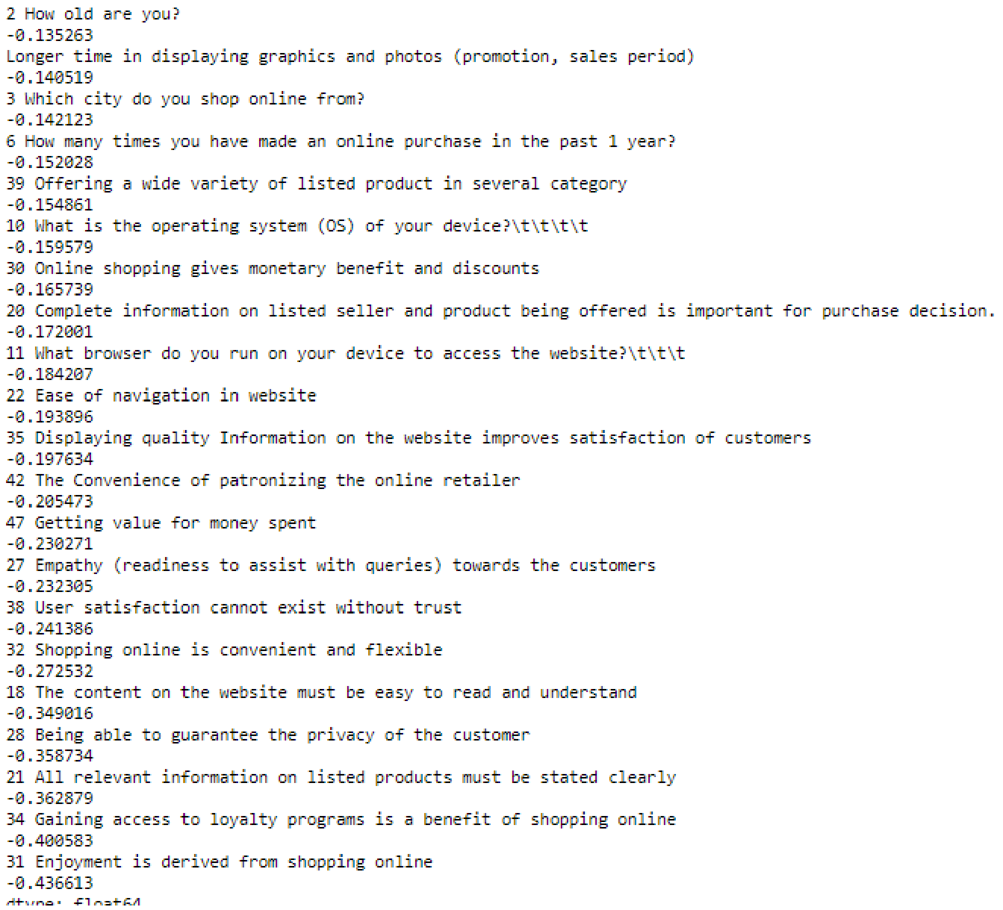
Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It’s a bi-variate analysis measure which describes the association between different variables. In most of the business it’s useful to express one subject in terms of its relationship with others.



```
corr_tar = customer.corrwith(customer['Which of the Indian online retailer would you recommend to a friend?'],axis=0)
pd.set_option('display.max_rows',None)
corr_tar.sort_values(ascending=False)
```

```
Which of the Indian online retailer would you recommend to a friend?
1.000000
Complete, relevant description information of products
0.680926
Reliability of the website or application
0.542711
Easy to use website or application
0.541713
Presence of online assistance through multi-channel
0.503836
Perceived Trustworthiness
0.483457
17 Why did you abandon the “Bag”, “Shopping Cart”?\t\t\t\t\t
0.448997
Longer delivery period
0.428419
Change in website/Application design
0.423877
Availability of several payment options
0.416729
Quickness to complete purchase
0.398754
Fast loading website speed of website and application
0.335192
Visual appealing web-page layout
0.316054
33 Return and replacement policy of the e-tailer is important for purchase decision
0.311562
15 What is your preferred payment Option?\t\t\t\t\t
0.308523
23 Loading and processing speed
0.298070
14 How much time do you explore the e- retail store before making a purchase decision?
0.290108
13 After first visit, how do you reach the online retail store?\t\t\t\t
0.279474
Longer page loading time (promotion, sales period)
0.278281
Longer time to get logged in (promotion, sales period)
0.261774
Website is as efficient as before
0.252154
40 Provision of complete and relevant product information
0.252121
Late declaration of price (promotion, sales period)
0.231029
Wild variety of product on offer
0.208213
```

```
45 You feel gratification shopping on your favorite e-tailer
0.179228
From the following, tick any (or all) of the online retailers you have shopped from;
0.170697
5 Since How Long You are Shopping Online ?
0.136106
19 Information on similar product to the one highlighted  is important for product comparison
0.127227
Frequent disruption when moving from one page to another
0.122953
16 How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart?\t\t\t\t\t\t\t
0.119196
8 Which device do you use to access the online shopping?
0.099425
26 Trust that the online retail store will fulfill its part of the transaction at the stipulated time
0.095426
44 Shopping on your preferred e-tailer enhances your social status
0.074666
9 What is the screen size of your mobile device?\t\t\t\t\t\t
0.074453
46 Shopping on the website helps you fulfill certain roles
0.069104
7 How do you access the internet while shopping on-line?
0.041129
29 Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)
0.035519
Limited mode of payment on most products (promotion, sales period)
0.028901
43 Shopping on the website gives you the sense of adventure
0.008540
1Gender of respondent
-0.003372
Security of customer financial information
-0.014895
36 User derive satisfaction while shopping on a good quality website or application
-0.024465
24 User friendly Interface of the website
-0.032348
25 Convenient Payment methods
-0.064096
12 Which channel did you follow to arrive at your favorite online store for the first time?
-0.071146
Privacy of customers’ information
-0.071876
41 Monetary savings
-0.079458
Speedy order delivery
-0.089890
4 What is the Pin Code of where you shop online from?
-0.097320
37 Net Benefit derived from shopping online can lead to users satisfaction
-0.126779
```



```
2 How old are you?
-0.135263
Longer time in displaying graphics and photos (promotion, sales period)
-0.140519
3 Which city do you shop online from?
-0.142123
6 How many times you have made an online purchase in the past 1 year?
-0.152028
39 Offering a wide variety of listed product in several category
-0.154861
10 What is the operating system (OS) of your device?\t\t\t\t
-0.159579
30 Online shopping gives monetary benefit and discounts
-0.165739
20 Complete information on listed seller and product being offered is important for purchase decision.
-0.172001
11 What browser do you run on your device to access the website?\t\t\t
-0.184207
22 Ease of navigation in website
-0.193896
35 Displaying quality Information on the website improves satisfaction of customers
-0.197634
42 The Convenience of patronizing the online retailer
-0.205473
47 Getting value for money spent
-0.230271
27 Empathy (readiness to assist with queries) towards the customers
-0.232305
38 User satisfaction cannot exist without trust
-0.241386
32 Shopping online is convenient and flexible
-0.272532
18 The content on the website must be easy to read and understand
-0.349016
28 Being able to guarantee the privacy of the customer
-0.358734
21 All relevant information on listed products must be stated clearly
-0.362879
34 Gaining access to loyalty programs is a benefit of shopping online
-0.400583
31 Enjoyment is derived from shopping online
-0.436613
dtype: float64
```

Feature Scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Standardization

The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. In case of multivariate data, this is done feature-wise (in other words independently for each column of the data).

Standardization:

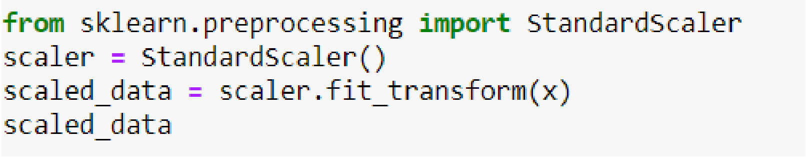
$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$



```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_data = scaler.fit_transform(x)
scaled_data
```

Principal Component Analysis

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the **Principal Components**. It is one of the popular tools that is used for exploratory data analysis and predictive modelling. It is a technique to draw strong patterns from the given dataset by reducing the variances.

PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are *image processing, movie recommendation system, optimizing the power allocation in various communication channels.* It is a feature extraction technique, so it contains the important variables and drops the least important variable.

The PCA algorithm is based on some mathematical concepts such as:

- Variance and Covariance
- Eigenvalues and Eigen factors

Some common terms used in PCA algorithm:

- **Dimensionality:** It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.

- **Correlation:** It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.
- **Orthogonal:** It defines that variable are not correlated to each other, and hence the correlation between the pair of variables is zero.
- **Eigenvectors:** If there is a square matrix M, and a non-zero vector v is given. Then v will be eigenvector if Av is the scalar multiple of v.
- **Covariance Matrix:** A matrix containing the covariance between the pair of variables is called the Covariance Matrix.

Principal Components in PCA

As described above, the transformed new features or the output of PCA are the Principal Components. The number of these PCs are either equal to or less than the original features present in the dataset. Some properties of these principal components are given below:

- The principal component must be the linear combination of the original features.
- These components are orthogonal, i.e., the correlation between a pair of variables is zero.
- The importance of each component decreases when going to 1 to n, it means the 1 PC has the most importance, and n PC will have the least importance.

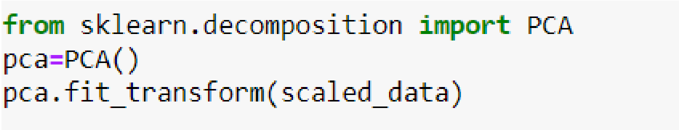
Steps for PCA algorithm

1. **Getting the dataset**
   Firstly, we need to take the input dataset and divide it into two subparts X and Y, where X is the training set, and Y is the validation set.
2. **Representing data into a structure**
   Now we will represent our dataset into a structure. Such as we will represent the two-dimensional matrix of independent variable X. Here each row corresponds to the data items, and the column corresponds to the Features. The number of columns is the dimensions of the dataset.
3. **Standardizing the data**
   In this step, we will standardize our dataset. Such as in a particular column, the features with high variance are more important compared to the features with lower variance.
   If the importance of features is independent of the variance of the feature, then we will divide each data item in a column with the standard deviation of the column. Here we will name the matrix as Z.

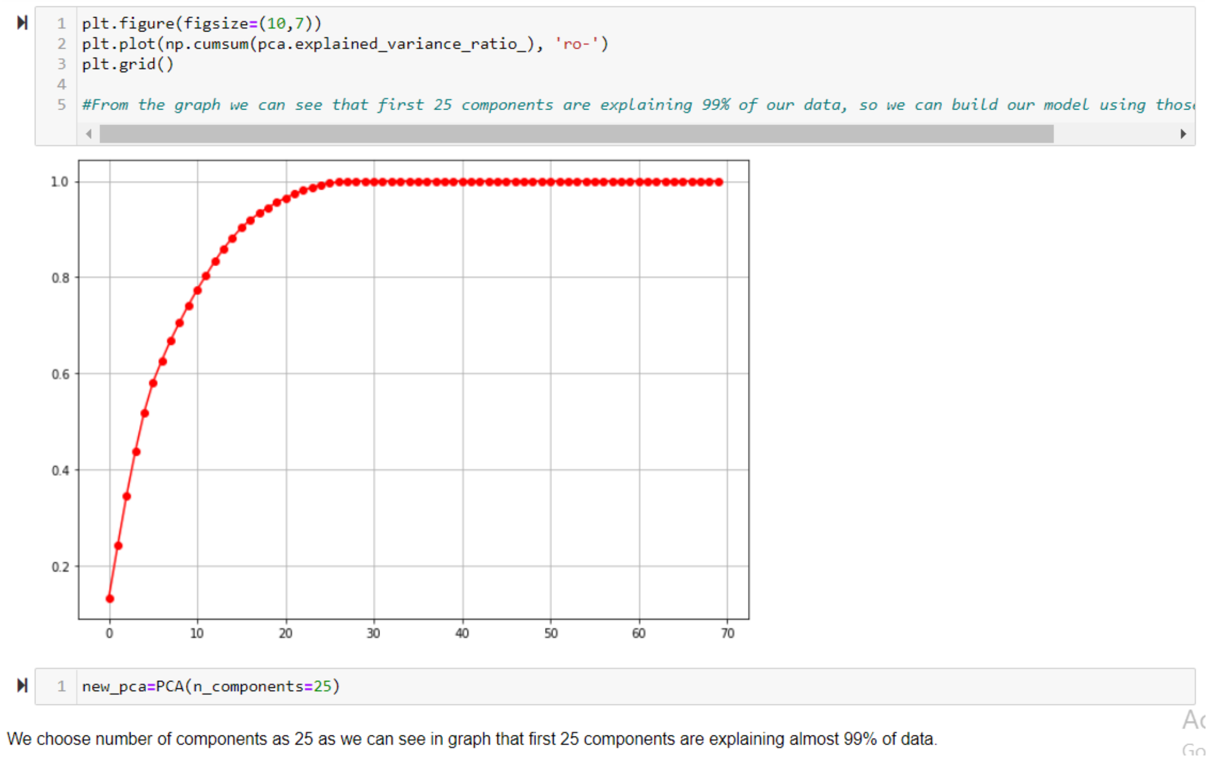
4. **Calculating the Covariance of Z**
   To calculate the covariance of Z, we will take the matrix Z, and will transpose it. After transpose, we will multiply it by Z. The output matrix will be the Covariance matrix of Z.
5. **Calculating the Eigen Values and Eigen Vectors**
   Now we need to calculate the eigenvalues and eigenvectors for the resultant covariance matrix Z. Eigenvectors or the covariance matrix are the directions of the axes with high information. And the coefficients of these eigenvectors are defined as the eigenvalues.
6. **Sorting the Eigen Vectors**
   In this step, we will take all the eigenvalues and will sort them in decreasing order, which means from largest to smallest. And simultaneously sort the eigenvectors accordingly in matrix P of eigenvalues. The resultant matrix will be named as P*.
7. **Calculating the new features Or Principal Components**
   Here we will calculate the new features. To do this, we will multiply the P* matrix to the Z. In the resultant matrix Z*, each observation is the linear combination of original features. Each column of the Z* matrix is independent of each other.
8. **Remove less or unimportant features from the new dataset.**
   The new feature set has occurred, so we will decide here what to keep and what to remove. It means, we will only keep the relevant or important features in the new dataset, and unimportant features will be removed out.

Applications of Principal Component Analysis

- PCA is mainly used as the dimensionality reduction technique in various AI applications such **as computer vision, image compression, etc.**
- It can also be used for finding hidden patterns if data has high dimensions. Some fields where PCA is used are Finance, data mining, Psychology, etc.



```
from sklearn.decomposition import PCA
pca=PCA()
pca.fit_transform(scaled_data)
```



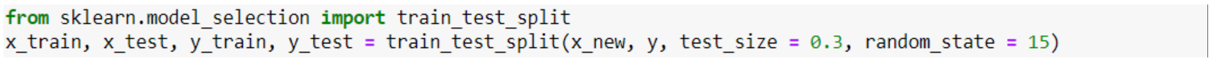
```
plt.figure(figsize=(10,7))
plt.plot(np.cumsum(pca.explained_variance_ratio_), 'ro-')
plt.grid()

#From the graph we can see that first 25 components are explaining 99% of our data, so we can build our model using thos
```

```
new_pca=PCA(n_components=25)
```

We choose number of components as 25 as we can see in graph that first 25 components are explaining almost 99% of data.

Model Building

A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfil its purpose.



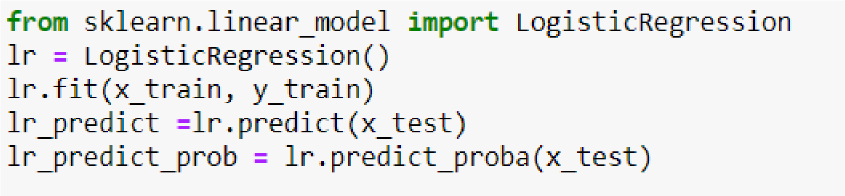
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_new, y, test_size = 0.3, random_state = 15)
```

Logistic regression

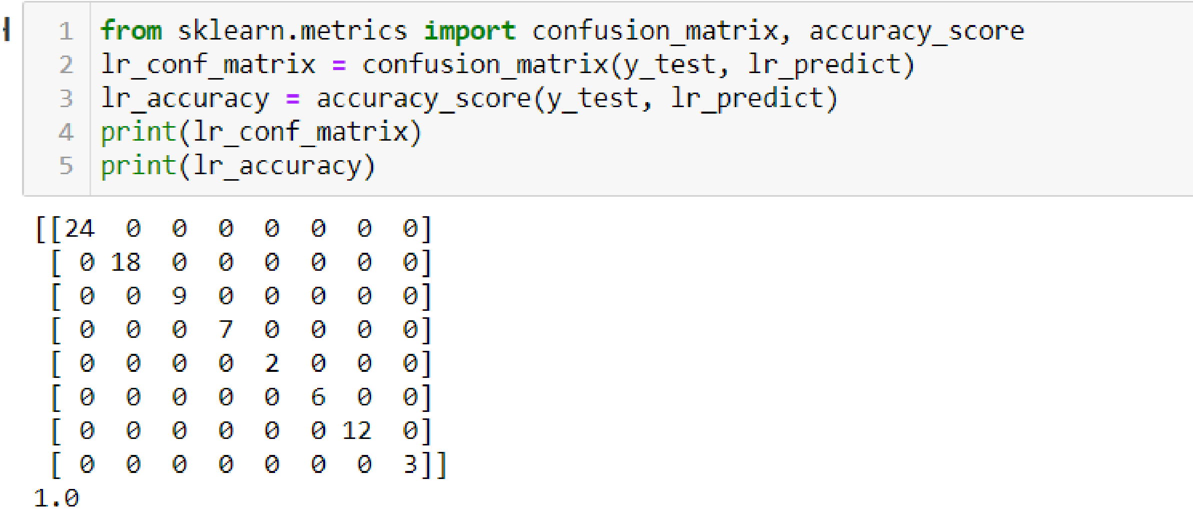
Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. As more relevant data comes in, the algorithm should get better at predicting classifications within data sets. Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis.

A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college.

The resulting analytical model can take into consideration multiple input criteria. In the case of college acceptance, the model could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.



```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(x_train, y_train)
lr_predict =lr.predict(x_test)
lr_predict_prob = lr.predict_proba(x_test)
```



```
from sklearn.metrics import confusion_matrix, accuracy_score
lr_conf_matrix = confusion_matrix(y_test, lr_predict)
lr_accuracy = accuracy_score(y_test, lr_predict)
print(lr_conf_matrix)
print(lr_accuracy)
```

```
[[24  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0]
 [ 0  0  9  0  0  0  0  0]
 [ 0  0  0  7  0  0  0  0]
 [ 0  0  0  0  2  0  0  0]
 [ 0  0  0  0  0  6  0  0]
 [ 0  0  0  0  0  0 12  0]
 [ 0  0  0  0  0  0  0  3]]
1.0
```

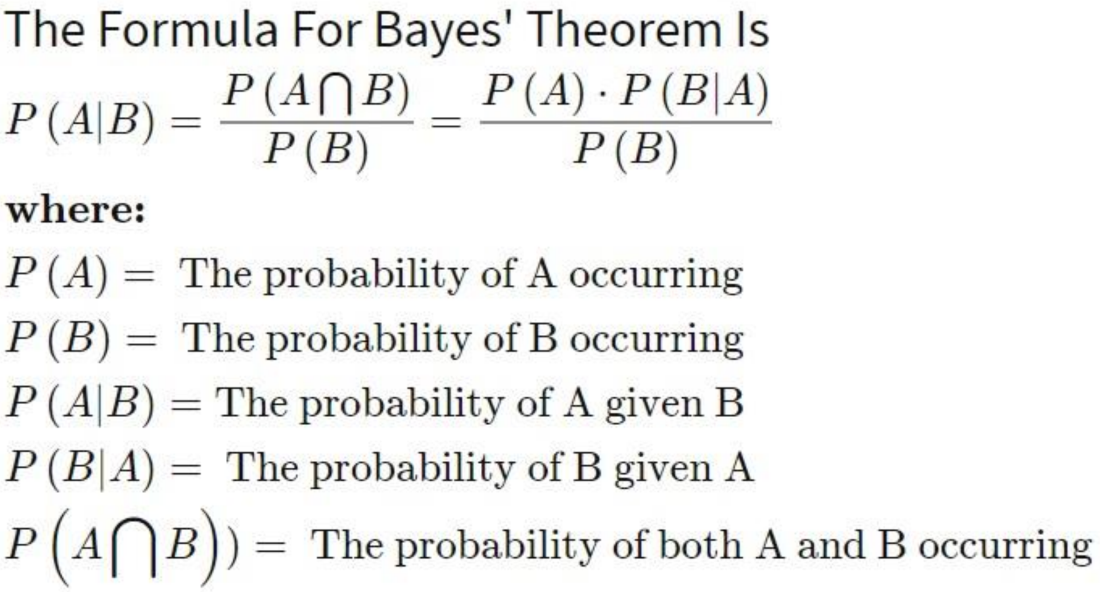
Gaussian Naive Bayes

Gaussian naive bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.

Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier.

Bayes Theorem

Bayes Theorem can be used to calculate conditional probability. Being a powerful tool in the study of probability, it is also applied in Machine Learning.



The Formula For Bayes' Theorem Is

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

where:

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

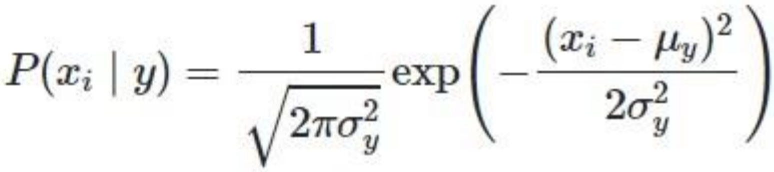
$P(B|A) =$ The probability of B given A

$P\left(A \bigcap B\right)) =$ The probability of both A and B occurring

Naive Bayes Classifier

Naive Bayes Classifiers are based on the Bayes Theorem. One assumption taken is the strong independence assumptions between the features. These classifiers assume that the value of a particular feature is independent of the value of any other feature. In a supervised learning situation, Naive Bayes Classifiers are trained very efficiently. Naive Bayed classifiers need a small training data to estimate the parameters needed for classification. Naive Bayes Classifiers have simple design and implementation and they can apply to many real-life situations.

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be-



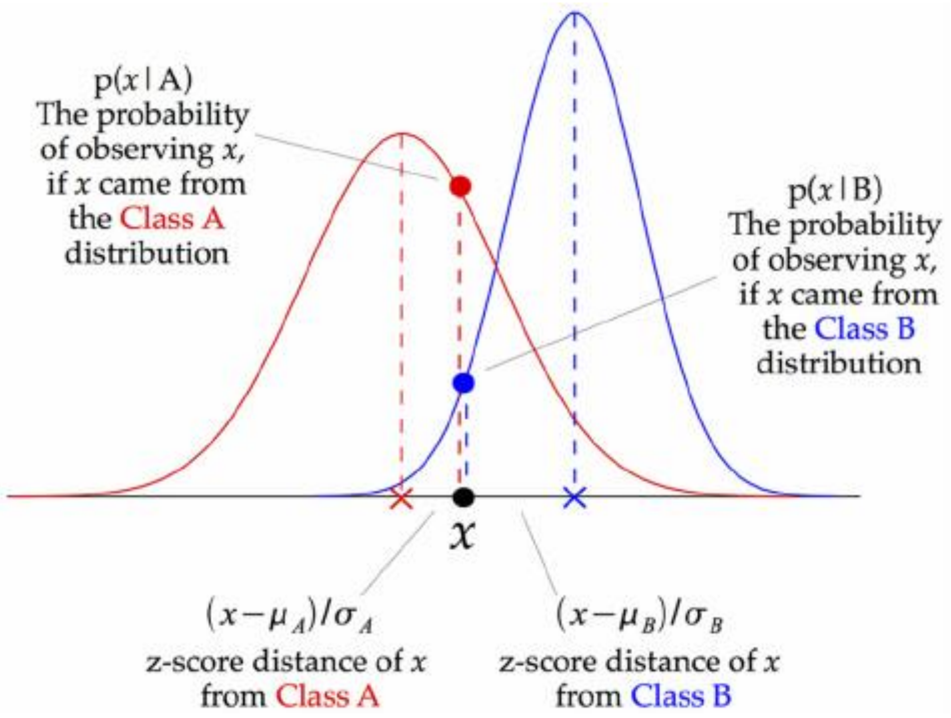
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Sometimes assume variance

- ✓ is independent of Y (i.e., σi),
- ✓ or independent of Xi (i.e., σk)
- ✓ or both (i.e., σ)

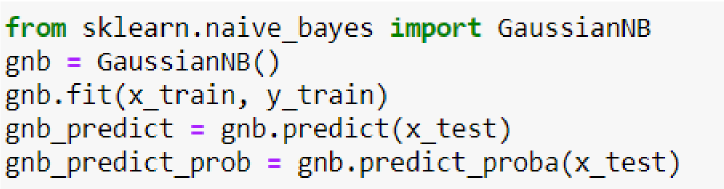
Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution.

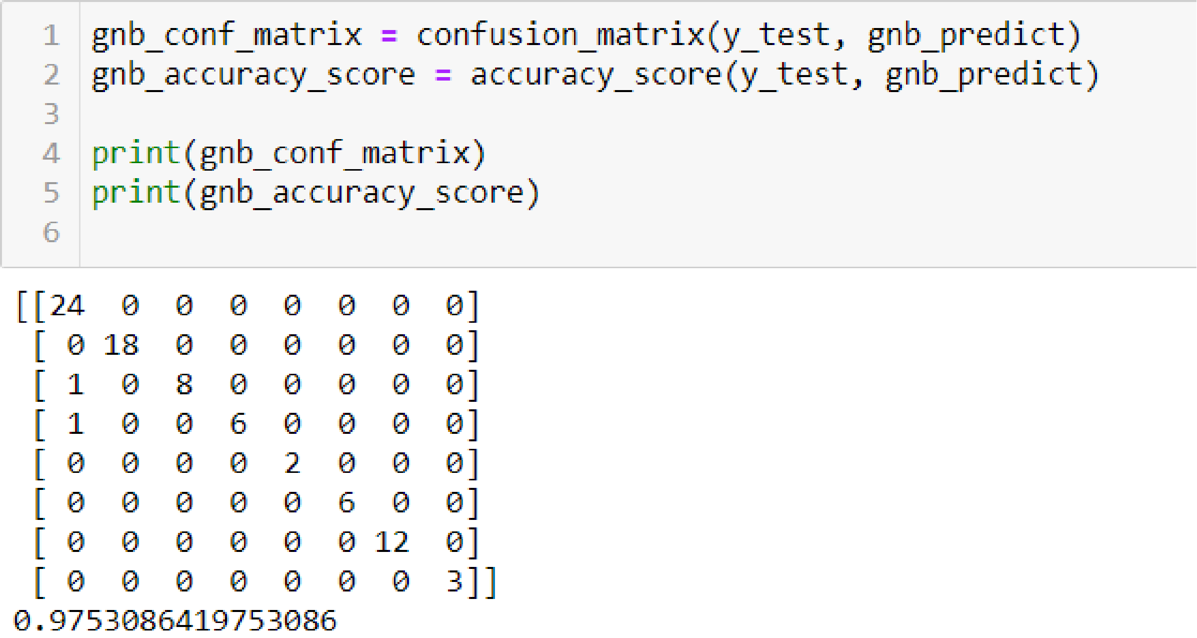

p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

The above illustration indicates how a Gaussian Naive Bayes (GNB) classifier works. At every data point, the z-score distance between that point and each class-mean is calculated, namely the distance from the class mean divided by the standard deviation of that class.

Thus, we see that the Gaussian Naive Bayes has a slightly different approach and can be used efficiently.



```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(x_train, y_train)
gnb_predict = gnb.predict(x_test)
gnb_predict_prob = gnb.predict_proba(x_test)
```



```
gnb_conf_matrix = confusion_matrix(y_test, gnb_predict)
gnb_accuracy_score = accuracy_score(y_test, gnb_predict)

print(gnb_conf_matrix)
print(gnb_accuracy_score)

```

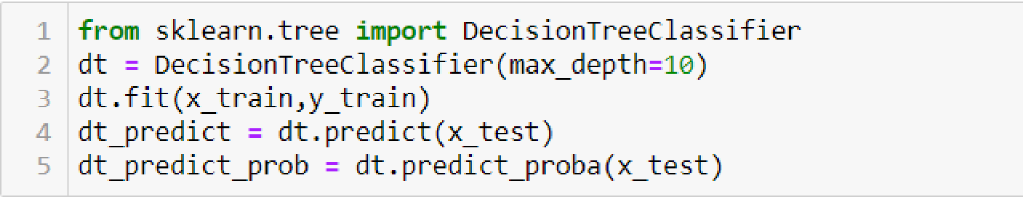
```
[[24  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0]
 [ 1  0  8  0  0  0  0  0]
 [ 1  0  0  6  0  0  0  0]
 [ 0  0  0  0  2  0  0  0]
 [ 0  0  0  0  0  6  0  0]
 [ 0  0  0  0  0  0 12  0]
 [ 0  0  0  0  0  0  0  3]]
0.9753086419753086
```

Decision Tree Algorithm

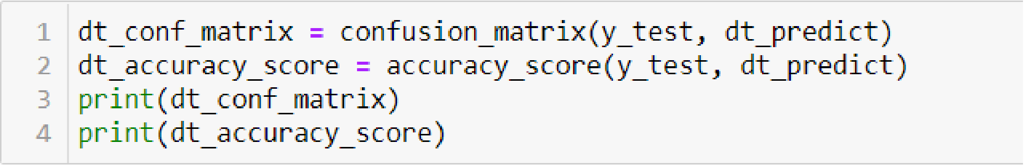
Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data (training data).

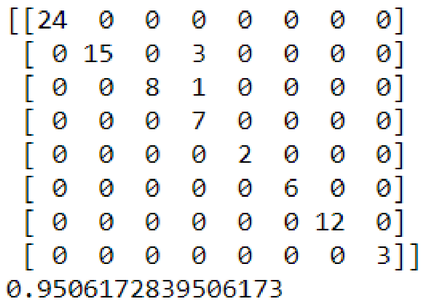
In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.



```python
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(max_depth=10)
dt.fit(x_train,y_train)
dt_predict = dt.predict(x_test)
dt_predict_prob = dt.predict_proba(x_test)
```



```python
dt_conf_matrix = confusion_matrix(y_test, dt_predict)
dt_accuracy_score = accuracy_score(y_test, dt_predict)
print(dt_conf_matrix)
print(dt_accuracy_score)
```



```
[[24  0  0  0  0  0  0  0]
 [ 0 15  0  3  0  0  0  0]
 [ 0  0  8  1  0  0  0  0]
 [ 0  0  0  7  0  0  0  0]
 [ 0  0  0  0  2  0  0  0]
 [ 0  0  0  0  0  6  0  0]
 [ 0  0  0  0  0  0 12  0]
 [ 0  0  0  0  0  0  0  3]]
0.9506172839506173
```

Random Forest

Random Forest is an example of ensemble learning, in which we combine multiple machine learning algorithms to obtain better predictive performance.

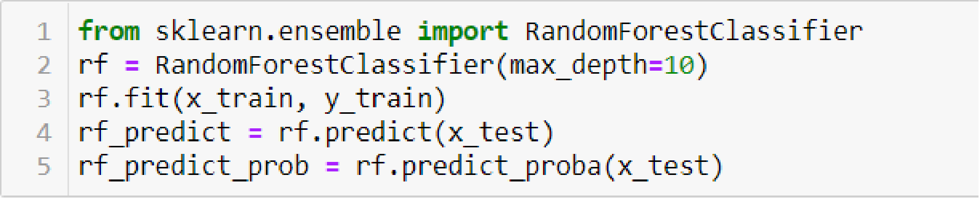
Why the name “Random”?

Two key concepts that give it the name random:

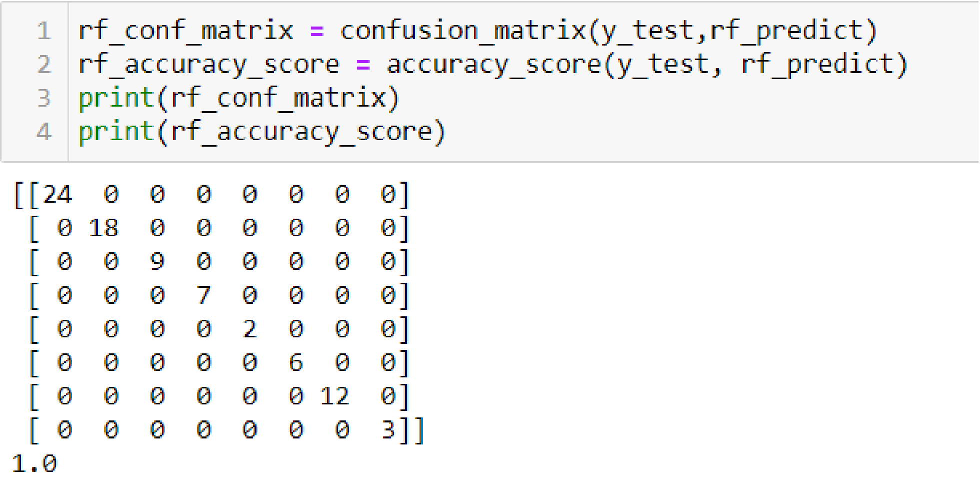
1. A random sampling of training data set when building trees.
2. Random subsets of features considered when splitting nodes.

A technique known as bagging is used to create an ensemble of trees where multiple training sets are generated with replacement.

In the bagging technique, a data set is divided into **N** samples using randomized sampling. Then, using a single learning algorithm a model is built on all samples. Later, the resultant predictions are combined using voting or averaging in parallel.



```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(max_depth=10)
rf.fit(x_train, y_train)
rf_predict = rf.predict(x_test)
rf_predict_prob = rf.predict_proba(x_test)
```

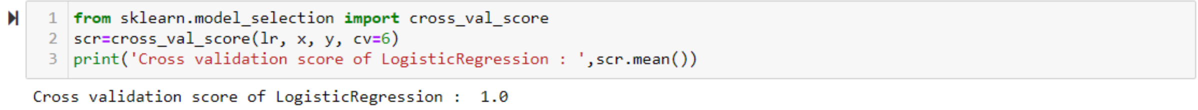


```
rf_conf_matrix = confusion_matrix(y_test,rf_predict)
rf_accuracy_score = accuracy_score(y_test, rf_predict)
print(rf_conf_matrix)
print(rf_accuracy_score)
```

```
[[24  0  0  0  0  0  0  0]
 [ 0 18  0  0  0  0  0  0]
 [ 0  0  9  0  0  0  0  0]
 [ 0  0  0  7  0  0  0  0]
 [ 0  0  0  0  2  0  0  0]
 [ 0  0  0  0  0  6  0  0]
 [ 0  0  0  0  0  0 12  0]
 [ 0  0  0  0  0  0  0  3]]
1.0
```

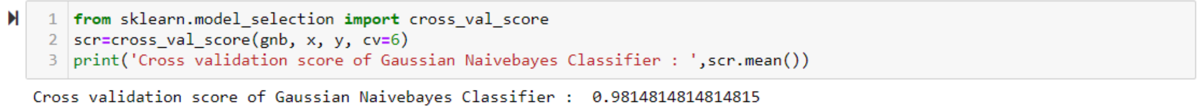
Cross Validation

The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).



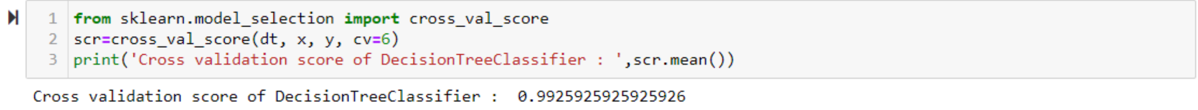
```
from sklearn.model_selection import cross_val_score
scr=cross_val_score(lr, x, y, cv=6)
print('Cross validation score of LogisticRegression : ',scr.mean())
```

```
Cross validation score of LogisticRegression :  1.0
```



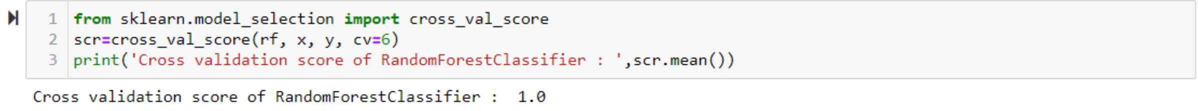
```
from sklearn.model_selection import cross_val_score
scr=cross_val_score(gnb, x, y, cv=6)
print('Cross validation score of Gaussian Naivebayes Classifier : ',scr.mean())
```

```
Cross validation score of Gaussian Naivebayes Classifier :  0.9814814814814815
```



```
from sklearn.model_selection import cross_val_score
scr=cross_val_score(dt, x, y, cv=6)
print('Cross validation score of DecisionTreeClassifier : ',scr.mean())
```

```
Cross validation score of DecisionTreeClassifier :  0.9925925925925926
```



```
from sklearn.model_selection import cross_val_score
scr=cross_val_score(rf, x, y, cv=6)
print('Cross validation score of RandomForestClassifier : ',scr.mean())
```

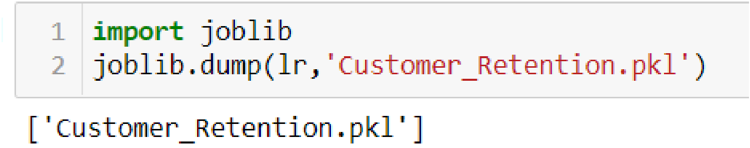
```
Cross validation score of RandomForestClassifier :  1.0
```

From above cross validation, we can observe that Logistic Regression and Random Forest Classifier models are having least difference between accuracy score and cross validation.

So, Logistic Regression Model or Random Forest Classifier with accuracy score of 100% can be as the best model

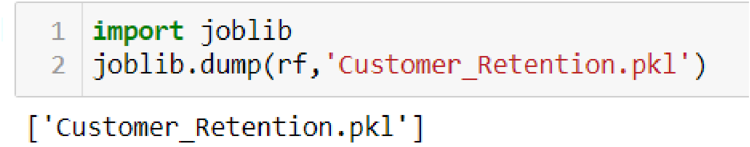
Deployment

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data.



```
import joblib
joblib.dump(lr,'Customer_Retention.pkl')
```

```
['Customer_Retention.pkl']
```



```
import joblib
joblib.dump(rf,'Customer_Retention.pkl')
```

```
['Customer_Retention.pkl']
```

Key Findings

- Most of them believe that amazon.in, flipkart.com will provide Complete, relevant description information of products and after them, people believe that amazon.in will provide complete information.
- Most of them believe that compared to all websites or applications, amazon.in is having more reliability of the website or application.
- Most of them believe that compared to all websites or applications, amazon.in is easy to use website or application.
- Most of them believe that amazon.in is having good presence of online assistance through multi-channel
- Most of them believe that amazon.in is having Perceived Trustworthiness when compared with other platforms
- Most of them believe that amazon.in is having longer delivery period when compared with other platforms
- Most of them believe that amazon.in is having frequent change in website/Application design when compared with other platforms
- Most of them believe that amazon.in, flipkart.com are having several payment options when compared with other platforms
- Most of them believe that amazon.com is quick in completing product purchase when compared with other platforms
- Most of them believe that amazon.in is having fastest loading website or application when compared with other platforms
- Most of them believe that amazon.in, flipkart.com are having visual appealing web-page layout when compared with other platforms.

- Most of them who shop in amazon.in believe that it is important to have return and replacement policy which effects the purchase decision.
- Most of them strongly agree that amazon.in and flipkart.com are having good loading and processing speed when compared with other platforms.
- Most of them believe that amazon.in, flipkart.com are having longer promotion, sales period when compared with other platforms.
- Most of them believe that amazon.in, flipkart.com are having longer promotion, sales period when compared with other platforms.
- Most of them believe that amazon.in is having website which is as efficient as before when compared with other platforms.

- Most of them believe that amazon.in is having provision of complete and relevant product information when compared with other platforms.
- Most of them believe that amazon.in, flipkart.com are having wild variety of product on offer when compared with other platforms.
- Most of them agree that shopping on amazon.in, flipkart.com feel gratification shopping when compared with other platforms.
- amazon.in, flipkart.com are having a greater number of customers who are shopping for more than 4 years.
- Most of them who shop in amazon.in, flipkart.com strongly agree that information on similar product to the one highlighted is important for product comparison.
- Most of them believe that amazon.in is having frequent disruption when moving from one page to another when compared with other platforms.
- Most of them use smartphones to access the online shopping.
- More number of people who are shopping on amazon.in, trust that the online retail store will fulfil its part of the transaction at the stipulated time.
- Most of them are having different opinion about shopping on their preferred e-tailer enhances your social status.
- Most of them agree that shopping on amazon.in, flipkart.com helps to fulfil certain roles.
- Most of them are using mobile internet to do shopping in online platforms.
- Most of them believe that amazon.in is having good responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.) when compared with other platforms.
- Most of them believe that snapdeal is having limited mode of payment on most products (promotion, sales period) when compared with other platforms.
- Most of them strongly agree that shopping on amazon.in website gives you the sense of adventure when compared with other platforms.
- We can see that female are spending more time on online shopping.
- Most of them believe that amazon.in, flipkart.com, snapdeal.com are having strong security of customer financial information when compared with other platforms.

- Most of them strongly agree that they derive satisfaction while shopping on a good quality website or application.
- Most of them who do online shopping on various platforms strongly agree that User friendly Interface of the website will be more effective for online shopping.
- Most of them who do online shopping on various platforms strongly agree that convenient payment methods will be more effective for online shopping.
- Most of them believe that amazon.in, flipkart.com are good at maintaining privacy of customers information when compared with other platforms.
- Most of them who do online shopping on various platforms strongly agree that monetary savings will affect online shopping.
- Most of them who do online shopping on various platforms strongly agree that net benefit derived from shopping online can lead to their satisfaction.
- Most of them believe that amazon.in, flipkart.com will take longer time in displaying graphics and photos (promotion, sales period) when compared with other platforms.
- Amazon.in is having a greater number of customers who shops for 31-40 times in a year.
- Amazon.in, Flipkart.com are having a greater number of customers who shops for more than 41 times in a year.
- Amazon.in is having a greater number of customers who shops for less than 10 times in a year.
- Most of them strongly agree that amazon.in is offering a wide variety of listed product in several category when compared with other platforms.
- Most of them strongly agree that online shopping gives monetary benefit and discounts
- Most of them strongly agree that online shopping gives monetary benefit and discounts
- Most of them strongly agree that ease of navigation in website will affect the customer reliability on website.
- Most of them strongly agree that displaying quality information on the website improves satisfaction of customers.
- Most of them agree that the convenience of patronizing the online retailer will affect the online shopping.
- Most of them strongly agree that getting value for money spent will affect the customer retention on their online platform.
- Most of them strongly agree that readiness to assist with queries in less time will affect the customer retention.
- Most of them strongly agree that user satisfaction cannot exist without trust.
- Most of them who shop online strongly agree that shopping online is convenient and flexible.
- Most of them strongly agree that the content on the website must be easy to read and understand.

- Most of them strongly agree that being able to guarantee the privacy of the customer will affect the customer retention.
- Most of them strongly agree that all relevant information on listed products must be stated clearly for ease in shopping.
- Most of them strongly agree that gaining access to loyalty programs is a benefit of shopping online.
- Most of them believe that amazon.in, flipkart.com are having longer promotion, sales period when compared with other platforms.

Conclusion:

As part of study, we can see huge potential for the online shopping in the future with the advancement in Information Technology infrastructure and awareness about the usage of internet in the rural as well as the urban areas, considering we successfully build a model following step by step procedure in model building.