FLIP ROBO

CUSTOMER RETENTION

Submitted by:

Gokula Krishnan R

Acknowledgement

I would like to express my gratitude to my guide Shubham Yadav (SME, Flip Robo) for his constant guidance, continuous encouragement, and unconditional help towards the development of the project. It was he who helped me whenever I got stuck somewhere in between. The project would have not been completed without his support and confidence he showed towards me.

Lastly, I would l like to thank all those who helped me directly or indirectly toward the successful completion of the project.

Introduction

Business Problem Framing

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.

As a Data Scientist, we have to apply our analytical skills to give findings and conclusions in detailed data analysis written in Jupyter notebook.

Conceptual Background of the Domain Problem

Customer Retention: Customer retention is a variety of activities aimed at keeping customers for the long term and turning them into loyal buyers. The end goal is transforming first-time customers into repeat customers and maximizing their lifetime value (LTV).

Importance of Customer Retention: Being mindful of customer retention matters because it helps you understand how loyal and satisfied your customers are, how strong your customer service is, and if there are any red flags that may turn off potential customers.

Focusing on customer retention pays dividends in the long run:

Lower cost compared to customer acquisition – As Econsultancy reports, 82% of companies state that customer retention is far cheaper than customer acquisition, yet many companies spend much more on acquisition instead of nurturing customers they already have.

Increased AOV – Not only is it much more cost-effective to retain current customers, but those shoppers are also willing to spend more as time goes on. Research shows that loyal customers are 23% more likely to spend with you than the average customer.

Increased profits – Taking care of customers and keeping them over time will also have a positive impact on your bottom line. Data shows that increasing customer retention by 5% can increase profits anywhere from 25% to 95% and that existing customers provide 65% of a company's business.

Brand ambassadors – The best thing about loyal customers is that they tend to share their positive experiences, thus becoming your brand ambassadors. That's priceless. According to Yotpo, 60% of consumers talk about a brand they're loyal to with their family and friends. As word of mouth increases exponentially, customer retention is a must-have for your business.

How to calculate customer retention rate: Customer retention rate (CRR) shows the percentage of customers that a company has retained over time.

To calculate your CRR, you should subtract the number of new customers acquired from the number of customers remaining at the end of the period. To calculate the percentage, divide that number by the total number of customers at the start and multiply by 100.

So, say you had 1,000 customers at the beginning of a quarter, acquired 200 new customers during the quarter and had 800 customers at the end of the quarter, your retention rate would be:

((800-200)/1,000)) x 100 = 60%

That means over the quarter you retained 60% of your existing customers.

Review of Literature

With the pandemic grabbing up places all over the world, the decline in businesses, macro and micro, is inevitable. Among these businesses is the retail sector which plays an important role for any economy. Retail sector covers all the basic necessities that a human being needs. Retail sector is dependent on many factors that drive the sales of a company and the most vital one is the customer retention. With eruption of technology, it is highly important to underscore the importance of social media and advanced technology. The main reason for that is the number of options the customers have. Due to this the switching cost is low. From the porter's analysis, we can clearly conclude that the bargaining power of the buyer is high for most retail commodities. Many customers are loyal towards a brand because of the brand image, so it is highly important that a company fulfils its CSR programs in a way that can grab the attention of the customers and glue them to the brand. This brand image can also be connected to the store ambience. This study is based on various factors that are perceived by the customers as of value and how these factors can enhance the customer retention.

Motivation for the Problem Undertaken

I am doing this for practice, to get more hands-on data exploration, Feature extraction and Model building.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

We would perform one type of supervised learning algorithms: regression. While it seems more reasonable to perform regression since house prices are continuous, we can also go for classification algorithm as classifying house prices into individual ranges of prices would also provide helpful insight for the Users. Also, this helps us explore different techniques which might be regression- or classification specific. Here, we will only perform classification. Since there are 71 features in the dataset, regularization is needed to prevent overfit. In order to determine the regularization parameter, throughout the project in regression part, we would first perform K-fold cross validation with k = 5 on a wide range of selection of regularization parameters; this helped us to select the best regularization parameters in the training phase. In order to further improve our models, we also performed principal component analysis pipeline on all models, and cross validated number of components to fit in each of the model to give the optimized results.

Data Sources and their formats

The data is collected from the Indian online shoppers. Results indicate the eretail success factors, which are very much critical for customer satisfaction.
There are two sheets (one is detailed) and second is encoded in the excel file.
You may use any of them by extracting in separate excel sheet. The number
of column(s) is more than 47.
The variables in this dataset are:
'1Gender of respondent',
'2 How old are you? ',
'3 Which city do you shop online from?',
'4 What is the Pin Code of where you shop online from?','5 Since How Long You are Shopping Online ?',
'6 How many times you have made an online purchase in the past 1 year?',
'7 How do you access the internet while shopping on-line?',
'8 Which device do you use to access the online shopping?',
'9 What is the screen size of your mobile device?\t\t\t\t\t\t
',
'10 What is the operating system (OS) of your device?\t\t\t\t
',
'11 What browser do you run on your device to access the website?\t\t\t
',
'12 Which channel did you follow to arrive at your favorite online store for

```
the first time? ',
'13 After first visit, how do you reach the online retail store?\t\t\t\t
',
'14 How much time do you explore the e- retail store before making a
purchase decision? ',
'15 What is your preferred payment Option?\t\t\t\t\t
',
'16 How frequently do you abandon (selecting an items and leaving without
making payment) your shopping cart?\t\t\t\t\t\t\t
',
'17 Why did you abandon the “Bag”, “Shopping Cart”?\t\t\t\t\t
',
'18 The content on the website must be easy to read and understand',
'19 Information on similar product to the one highlighted is important for
product comparison',
'20 Complete information on listed seller and product being offered is
important for purchase decision.',
'21 All relevant information on listed products must be stated clearly',
'22 Ease of navigation in website',
'23 Loading and processing speed',
'24 User friendly Interface of the website',
'25 Convenient Payment methods',
'26 Trust that the online retail store will fulfill its part of the transaction at
the stipulated time',
'27 Empathy (readiness to assist with queries) towards the customers',
'28 Being able to guarantee the privacy of the customer',
'29 Responsiveness, availability of several communication channels (email,
online rep, twitter, phone etc.)',
'30 Online shopping gives monetary benefit and discounts',
'31 Enjoyment is derived from shopping online',
'32 Shopping online is convenient and flexible',
'33 Return and replacement policy of the e-tailer is important for purchase
decision',
'34 Gaining access to loyalty programs is a benefit of shopping online',
'35 Displaying quality Information on the website improves satisfaction of
customers','36 User derive satisfaction while shopping on a good quality website or
application',
'37 Net Benefit derived from shopping online can lead to users satisfaction',
'38 User satisfaction cannot exist without trust',
'39 Offering a wide variety of listed product in several category',
'40 Provision of complete and relevant product information',
'41 Monetary savings',
'42 The Convenience of patronizing the online retailer',
'43 Shopping on the website gives you the sense of adventure',
```

'44 Shopping on your preferred e-tailer enhances your social status',
'45 You feel gratification shopping on your favorite e-tailer',
'46 Shopping on the website helps you fulfill certain roles',
'47 Getting value for money spent',
'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?'

There are 269 records and total 71 columns or variables, out of them 47 columns are information regarding the users and others are their views on experiencing the e-commerce websites

Data Pre-processing Done

I didn't find missing values in both train and test data set. Columns names are big and have lots of spaces in it. Replace all the spaces with __ So that I can able to call them. Checked for null vales. Didn't found any.

Hardware and Software Requirements and Tools Used

The system requirements are description of features and functionalities of the target system. Requirements convey the expectations of users from the software product. The requirements can be obvious or hidden, known or unknown, expected or unexpected from client's point of view. System requirements are all of the requirements at the system level that describe the functions which the system as a whole should fulfil to satisfy the stakeholder needs and requirements, and is expressed in an appropriate combination of textual statements, views, and non-functional requirements; the latter expressing the levels of safety, security, reliability, etc., that will be necessary.

Hardware requirements: -

1. Processor — core i5 and above
2. RAM — 8 GB or above
3. SSD — 250GB or above

Software requirements: -

Anaconda

Libraries: -

From sklearn.preprocessing import StandardScaler

As these columns are different in **scale**, they are **standardized** to have common **scale** while building machine learning model. This is useful when you want to compare data that correspond to different units.

from sklearn.preprocessing import Label Encoder

Label Encoder and One Hot Encoder. These two encoders are parts of the SciKit Learn library in Python, and they are used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

from sklearn.model_selection import train_test_split,cross_val_score

Train_test_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train_test_split will make random partitions for the two subsets.

The algorithm is trained and tested K times, each time a new set is used as testing set while remaining sets are used for training. Finally, the result of the K-Fold Cross-Validation is the average of the results obtained on each set.

from sklearn.neighbors import KNeighborsClassifier

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition

from sklearn.linear_model import LogisticRegression

The library sklearn can be used to perform logistic regression in a few lines as shown using the LogisticRegression class. It also supports multiple features. It requires the input values to be in a specific format hence they have been reshaped before training using the fit method.

from sklearn.tree import DecisionTreeClassifier

Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

For feature transformation I have used Log normal transformation to make the continuous non zero variables close to normal distributed. Use of Pearson’s correlation coefficient to check the correlation between dependent and independent features. Use of Min Max scaler to scale down the features and one label encoding to encode categorical features in numeric. Used PCA to decrease number of cloumns.

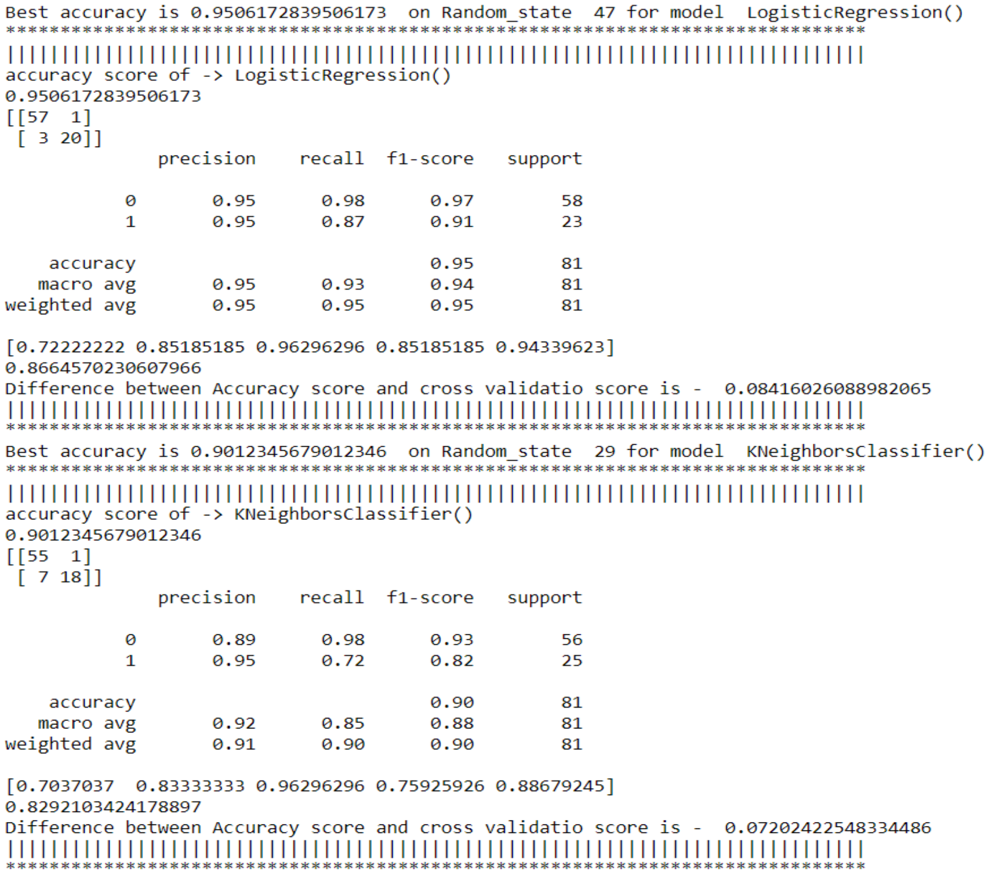
Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

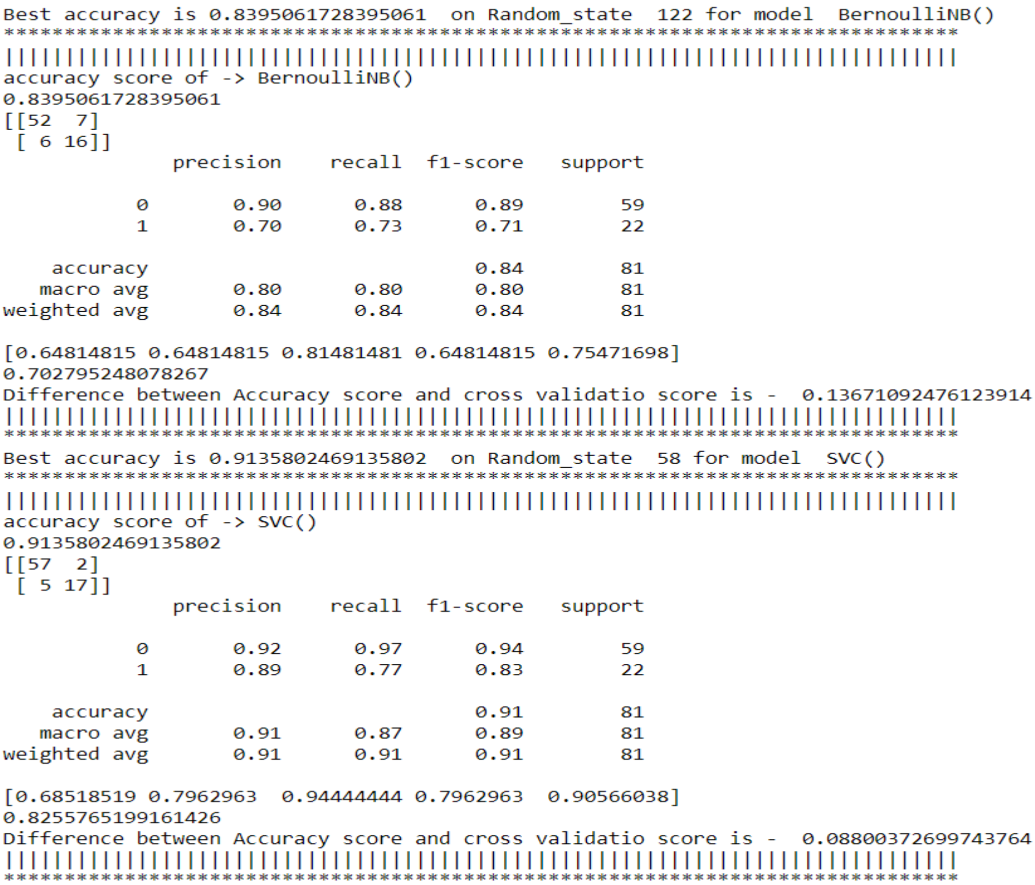
- KNN = KNeighborsClassifier()
- LR = LogisticRegression()
- BNB = BernoulliNB()
- DT = DecisionTreeClassifier()
- RF = RandomForestClassifier()

I applied all these algorithms in the dataset.

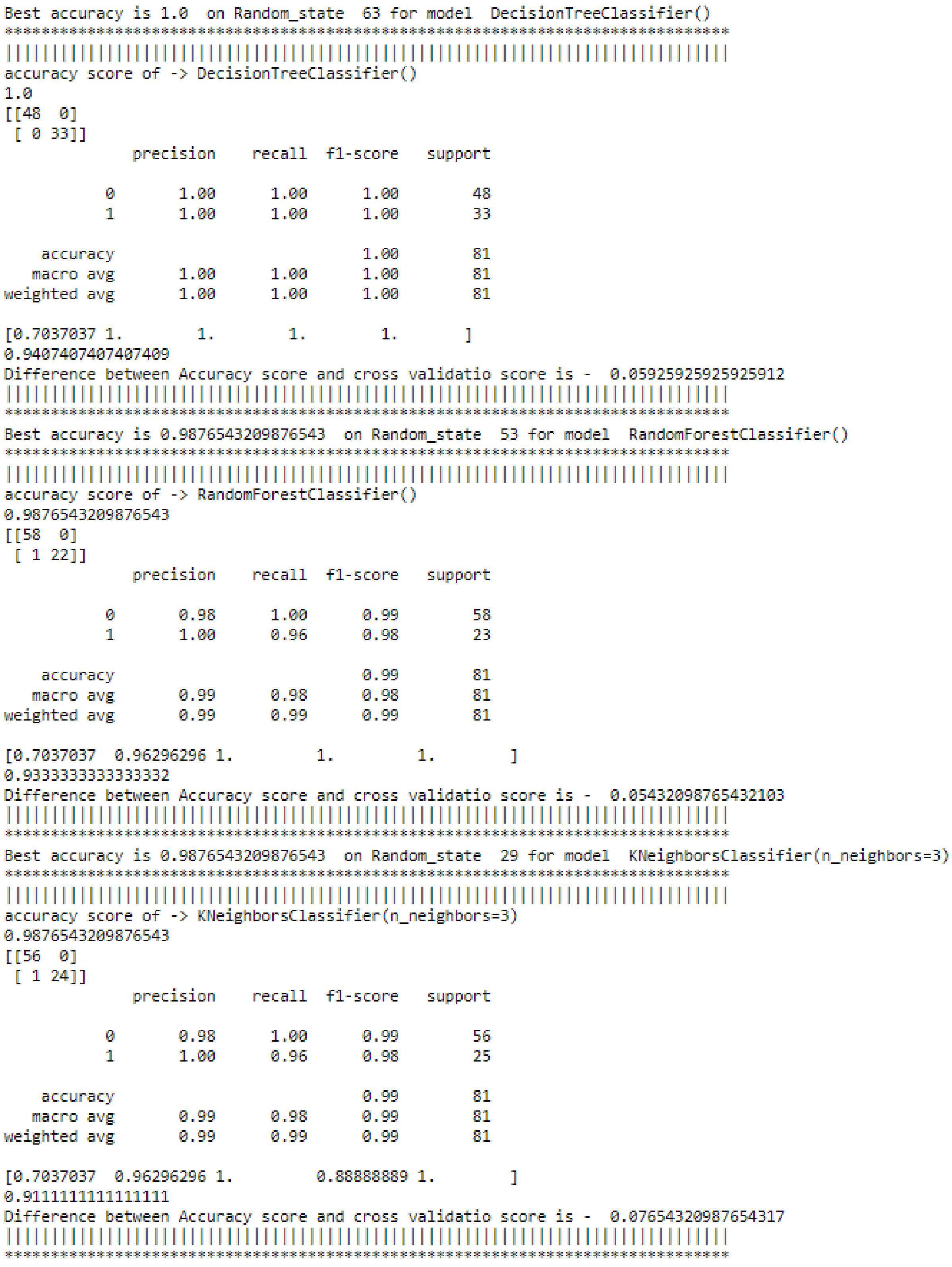
Run and Evaluate selected models



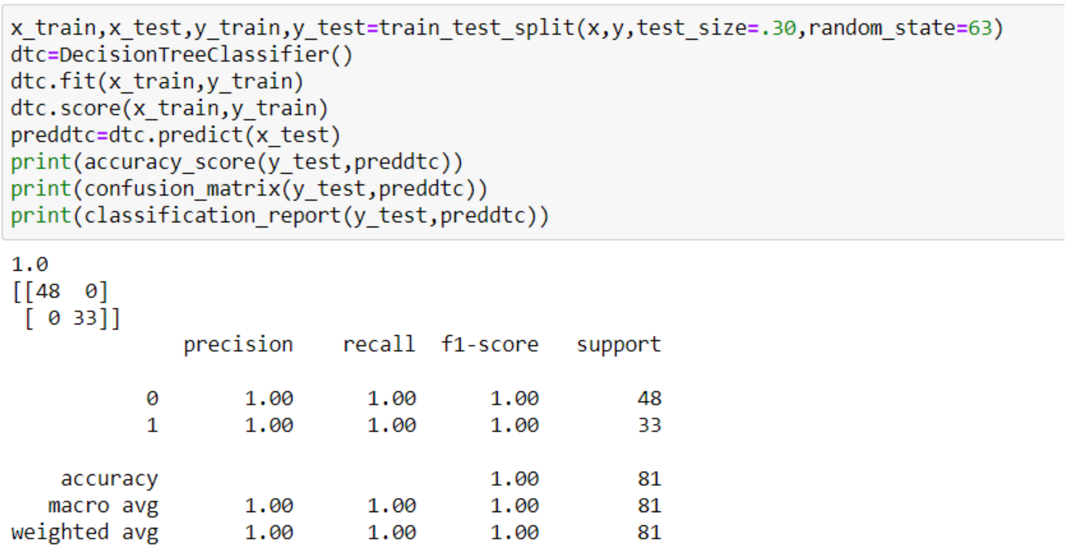
```
Best accuracy is 0.9506172839506173  on Random_state  47 for model  LogisticRegression()
******************************************************************************
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> LogisticRegression()
0.9506172839506173
[[57  1]
 [ 3 20]]
              precision    recall  f1-score   support

           0       0.95      0.98      0.97        58
           1       0.95      0.87      0.91        23

    accuracy                           0.95        81
   macro avg       0.95      0.93      0.94        81
weighted avg       0.95      0.95      0.95        81

[0.72222222 0.85185185 0.96296296 0.85185185 0.94339623]
0.8664570230607966
Difference between Accuracy score and cross validatio score is -  0.08416026088982065
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
******************************************************************************
Best accuracy is 0.9012345679012346  on Random_state  29 for model  KNeighborsClassifier()
******************************************************************************
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> KNeighborsClassifier()
0.9012345679012346
[[55  1]
 [ 7 18]]
              precision    recall  f1-score   support

           0       0.89      0.98      0.93        56
           1       0.95      0.72      0.82        25

    accuracy                           0.90        81
   macro avg       0.92      0.85      0.88        81
weighted avg       0.91      0.90      0.90        81

[0.7037037  0.83333333 0.96296296 0.75925926 0.88679245]
0.8292103424178897
Difference between Accuracy score and cross validatio score is -  0.07202422548334486
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
******************************************************************************
```



```
Best accuracy is 0.8395061728395061  on Random_state  122 for model  BernoulliNB()
******************************************************************************
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> BernoulliNB()
0.8395061728395061
[[52  7]
 [ 6 16]]
              precision    recall  f1-score   support

           0       0.90      0.88      0.89        59
           1       0.70      0.73      0.71        22

    accuracy                           0.84        81
   macro avg       0.80      0.80      0.80        81
weighted avg       0.84      0.84      0.84        81

[0.64814815 0.64814815 0.81481481 0.64814815 0.75471698]
0.702795248078267
Difference between Accuracy score and cross validatio score is -  0.13671092476123914
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
******************************************************************************
Best accuracy is 0.9135802469135802  on Random_state  58 for model  SVC()
******************************************************************************
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> SVC()
0.9135802469135802
[[57  2]
 [ 5 17]]
              precision    recall  f1-score   support

           0       0.92      0.97      0.94        59
           1       0.89      0.77      0.83        22

    accuracy                           0.91        81
   macro avg       0.91      0.87      0.89        81
weighted avg       0.91      0.91      0.91        81

[0.68518519 0.7962963  0.94444444 0.7962963  0.90566038]
0.8255765199161426
Difference between Accuracy score and cross validatio score is -  0.08800372699743764
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
******************************************************************************
```

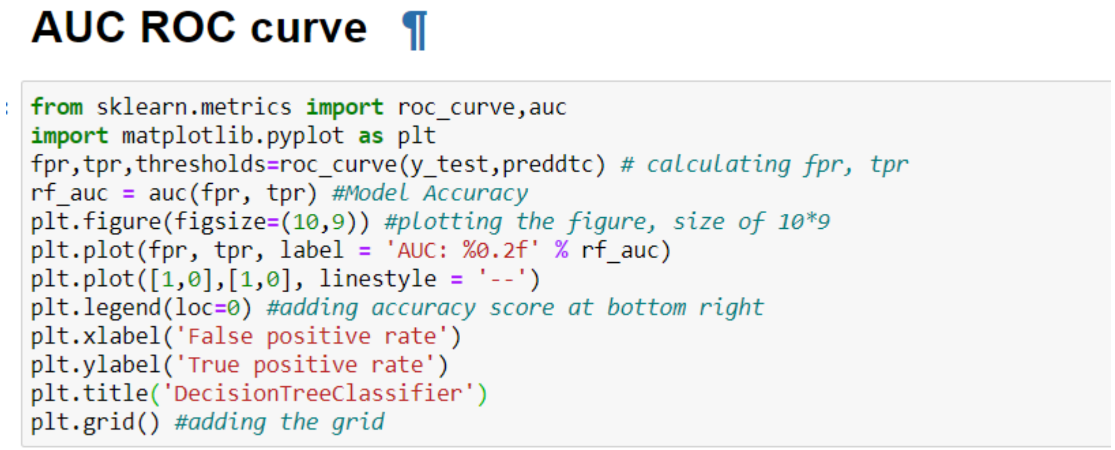


```
Best accuracy is 1.0  on Random_state  63 for model  DecisionTreeClassifier()
*******************************************************************************
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> DecisionTreeClassifier()
1.0
[[48  0]
 [ 0 33]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        48
           1       1.00      1.00      1.00        33

    accuracy                           1.00        81
   macro avg       1.00      1.00      1.00        81
weighted avg       1.00      1.00      1.00        81

[0.7037037 1.        1.        1.        1.       ]
0.9407407407407409
Difference between Accuracy score and cross validatio score is -  0.05925925925925912
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
*******************************************************************************
Best accuracy is 0.9876543209876543  on Random_state  53 for model  RandomForestClassifier()
*******************************************************************************
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> RandomForestClassifier()
0.9876543209876543
[[58  0]
 [ 1 22]]
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        58
           1       1.00      0.96      0.98        23

    accuracy                           0.99        81
   macro avg       0.99      0.98      0.98        81
weighted avg       0.99      0.99      0.99        81

[0.7037037  0.96296296 1.         1.         1.        ]
0.9333333333333332
Difference between Accuracy score and cross validatio score is -  0.05432098765432103
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
*******************************************************************************
Best accuracy is 0.9876543209876543  on Random_state  29 for model  KNeighborsClassifier(n_neighbors=3)
*******************************************************************************
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accuracy score of -> KNeighborsClassifier(n_neighbors=3)
0.9876543209876543
[[56  0]
 [ 1 24]]
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        56
           1       1.00      0.96      0.98        25

    accuracy                           0.99        81
   macro avg       0.99      0.98      0.99        81
weighted avg       0.99      0.99      0.99        81

[0.7037037  0.96296296 1.         0.88888889 1.        ]
0.9111111111111111
Difference between Accuracy score and cross validatio score is -  0.07654320987654317
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
*******************************************************************************
```



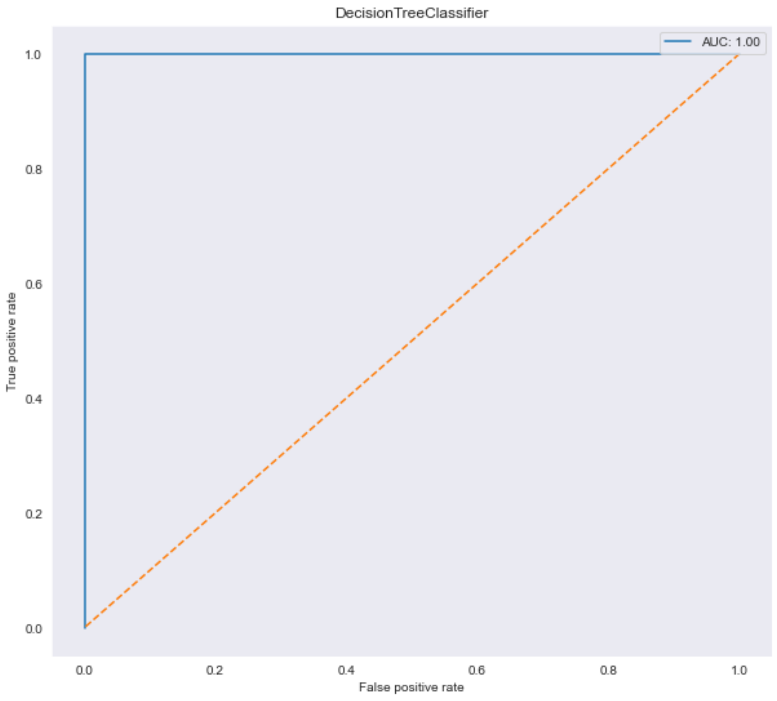
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.30,random_state=63)
dtc=DecisionTreeClassifier()
dtc.fit(x_train,y_train)
dtc.score(x_train,y_train)
preddtc=dtc.predict(x_test)
print(accuracy_score(y_test,preddtc))
print(confusion_matrix(y_test,preddtc))
print(classification_report(y_test,preddtc))
```

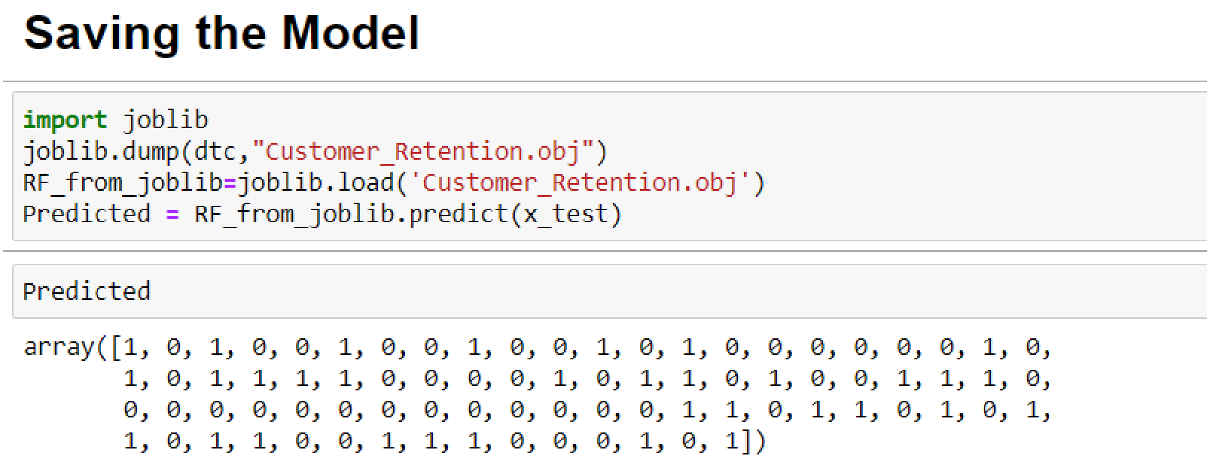
```
1.0
[[48  0]
 [ 0 33]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        48
           1       1.00      1.00      1.00        33

    accuracy                           1.00        81
   macro avg       1.00      1.00      1.00        81
weighted avg       1.00      1.00      1.00        81
```



AUC ROC curve ¶

```
from sklearn.metrics import roc_curve,auc
import matplotlib.pyplot as plt
fpr,tpr,thresholds=roc_curve(y_test,preddtc) # calculating fpr, tpr
rf_auc = auc(fpr, tpr) #Model Accuracy
plt.figure(figsize=(10,9)) #plotting the figure, size of 10*9
plt.plot(fpr, tpr, label = 'AUC: %0.2f' % rf_auc)
plt.plot([1,0],[1,0], linestyle = '--')
plt.legend(loc=0) #adding accuracy score at bottom right
plt.xlabel('False positive rate')
plt.ylabel('True positive rate')
plt.title('DecisionTreeClassifier')
plt.grid() #adding the grid
```





Saving the Model

```
import joblib
joblib.dump(dtc,"Customer_Retention.obj")
RF_from_joblib=joblib.load('Customer_Retention.obj')
Predicted = RF_from_joblib.predict(x_test)
```

```
Predicted
```

```
array([1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1,
       1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1])
```

Key Metrics for success in solving problem under consideration

Precision: can be seen as a measure of quality, **higher precision** means that an algorithm returns more relevant results than irrelevant ones.

Recall is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.

Accuracy score is used when the True Positives and True negatives are more important. **Accuracy** can be used when the class distribution is similar.

F1-score is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.

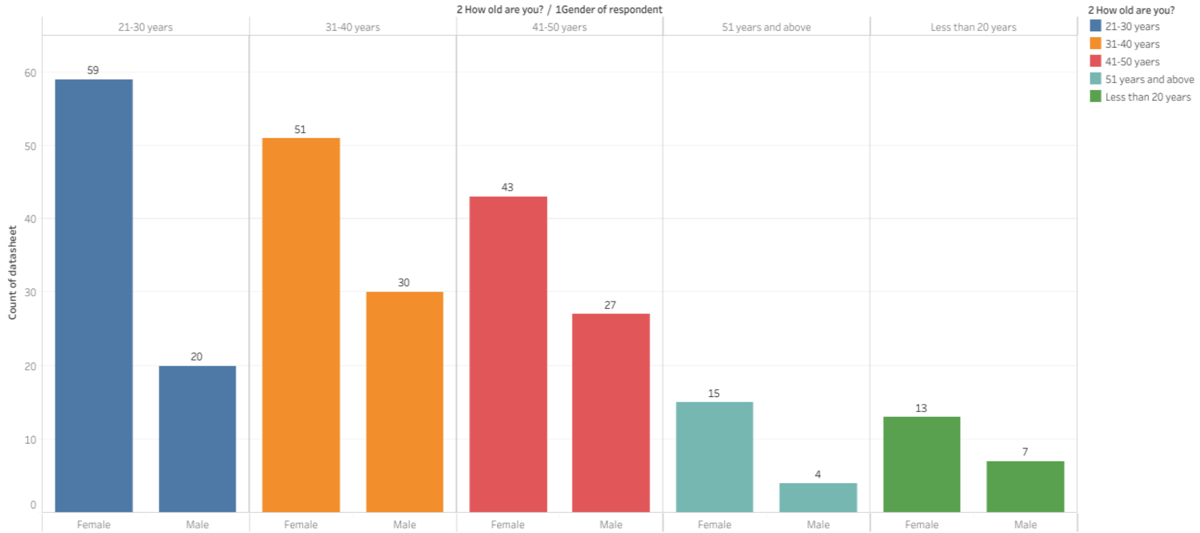
Cross_val_score :- To run **cross-validation** on multiple metrics and also to return train **scores**, fit times and **score** times. Get predictions from each split of **cross-validation** for diagnostic purposes. Make a scorer from a performance metric or loss function.

AUC_ROC _score :- ROC curve. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0

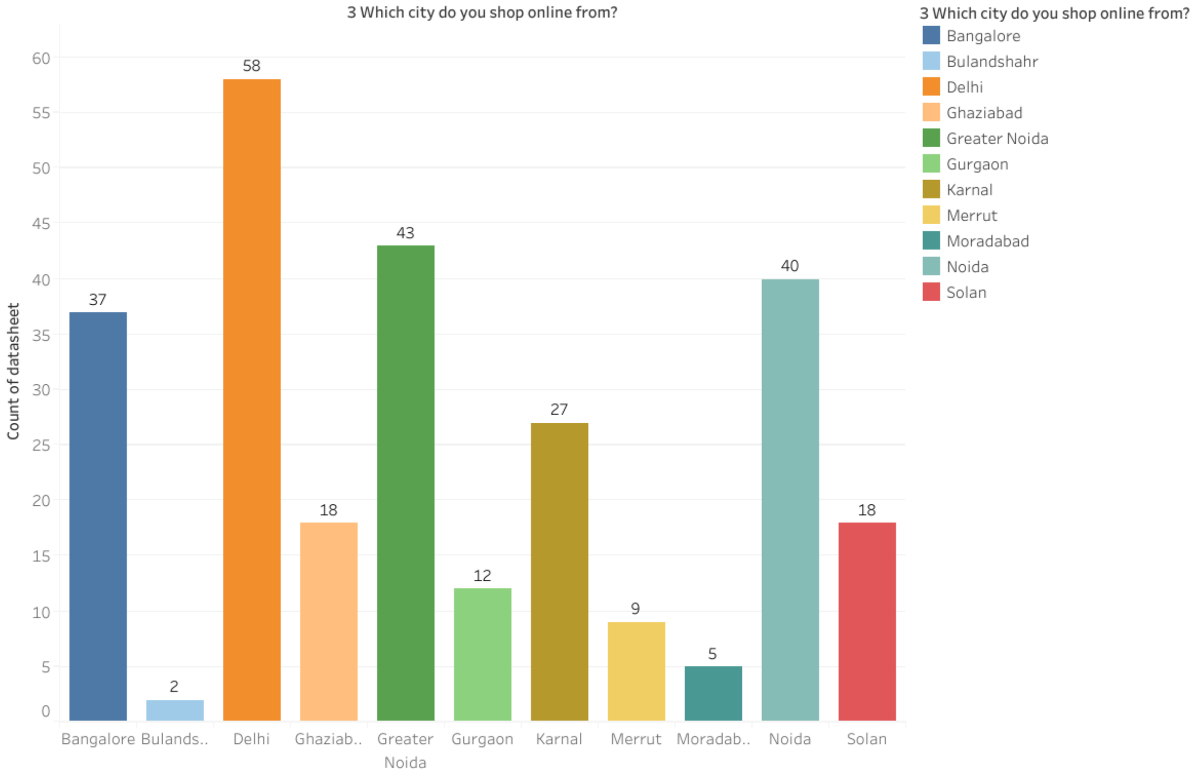
Visualizations

Finding1: So from the below graph we can see that, There are 269 users and among them 181 are females and 88 are male. We have five age groups i.e. Less than 20 years, 21-30 years, 31-40 years, 41-50 years & 51 years and above.

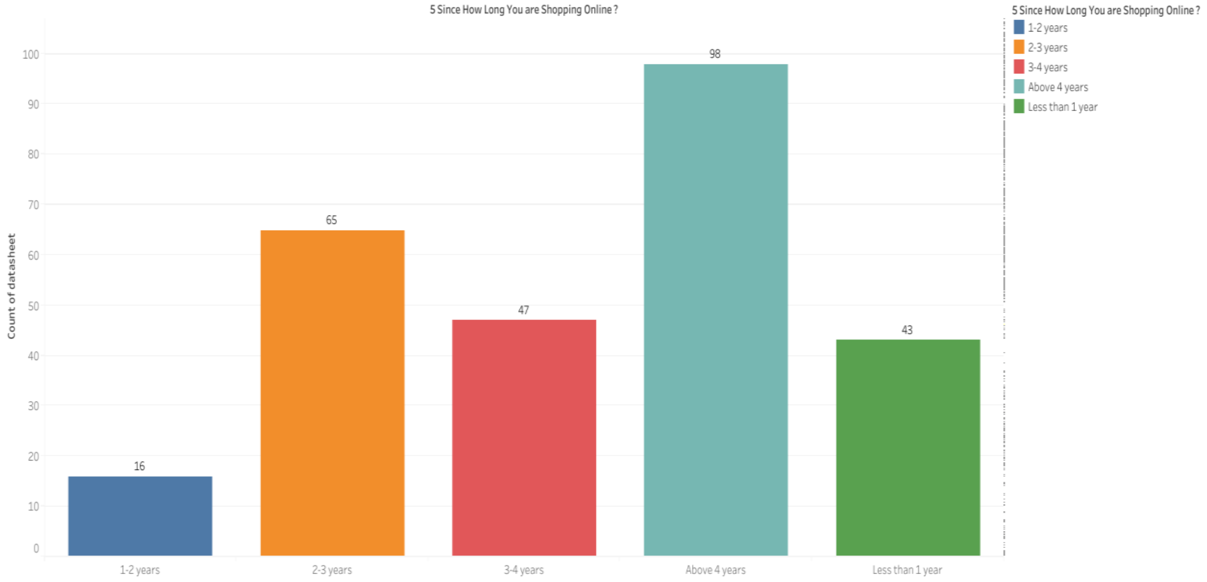
Among them age in group of 31-40 years there are 81 users, 79 in the age group of 21-30 years, 70 in the group of 41-50 years and and 19 users are above 51 years.



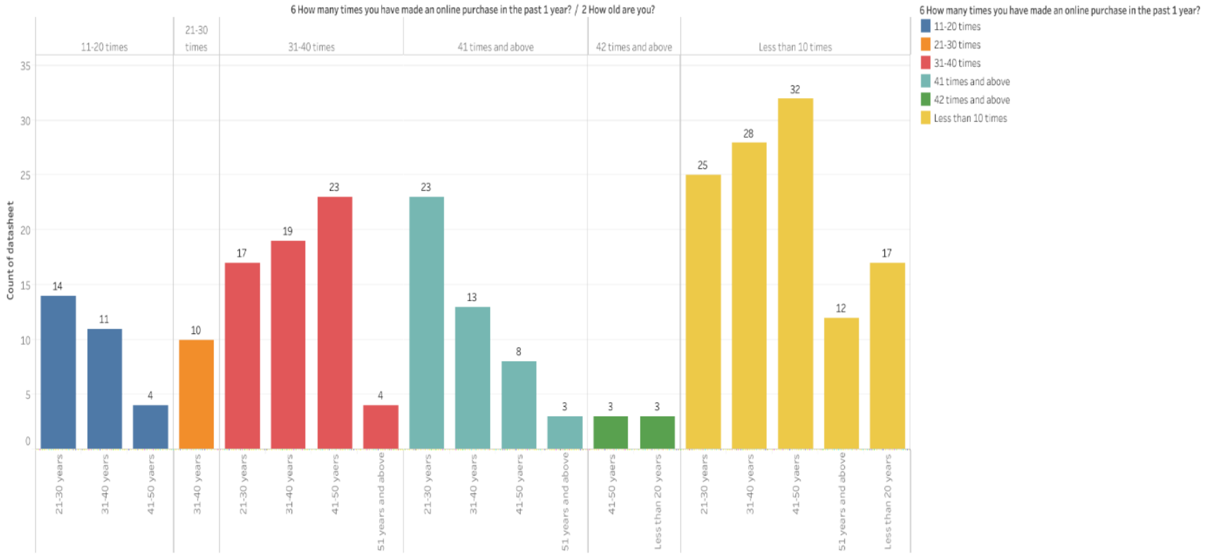
Finding 2: The all 269 users are taken from 11 different cities, that are Bangalore, Bulandsahar, Delhi, Ghaziabad, Greater Noida, Gurgaon, Karnal, Meerut, Moradabad, Noida, Solan. Maximum 58 uses are from Delhi and minimum 2 are from Bulandsahar.



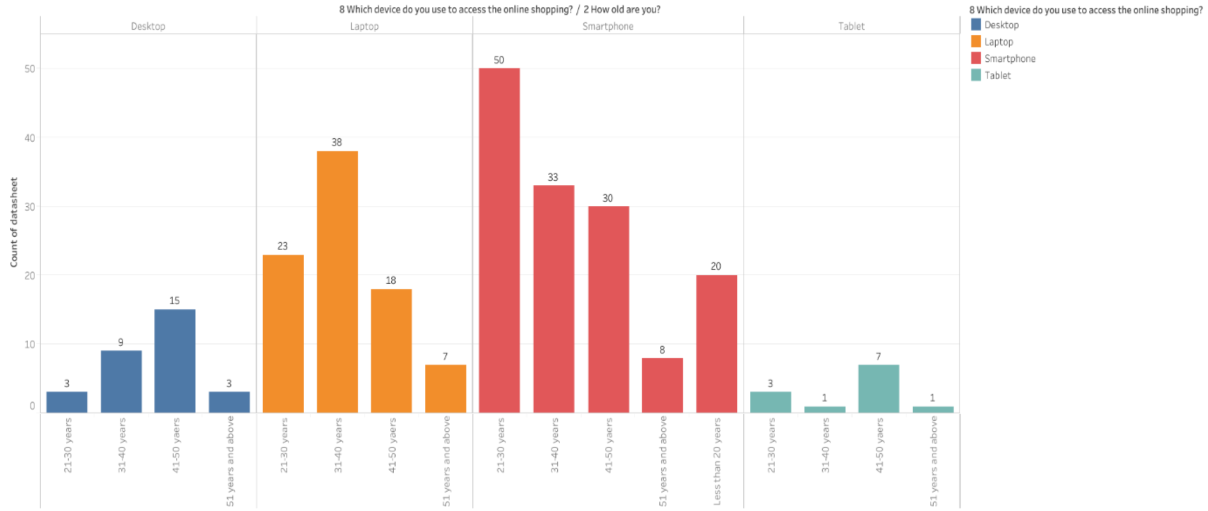
Finding 3: From our customers 171 customers started online shopping within last years 4 years and 98 have started the same for above 4 years, among them 36 customers are in age-group 31-40years and 30 are in 21-30years age-group.



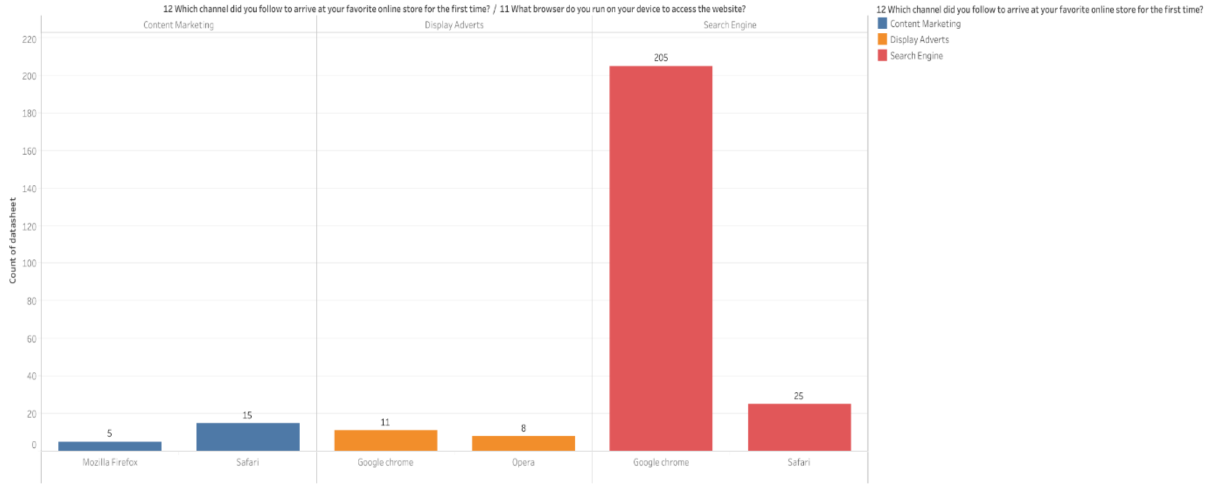
Finding 4: Here our finding is, the age-group 21-30years have made more no. times of online purchases than others in last year foloowed by the age-group 31-40 years.



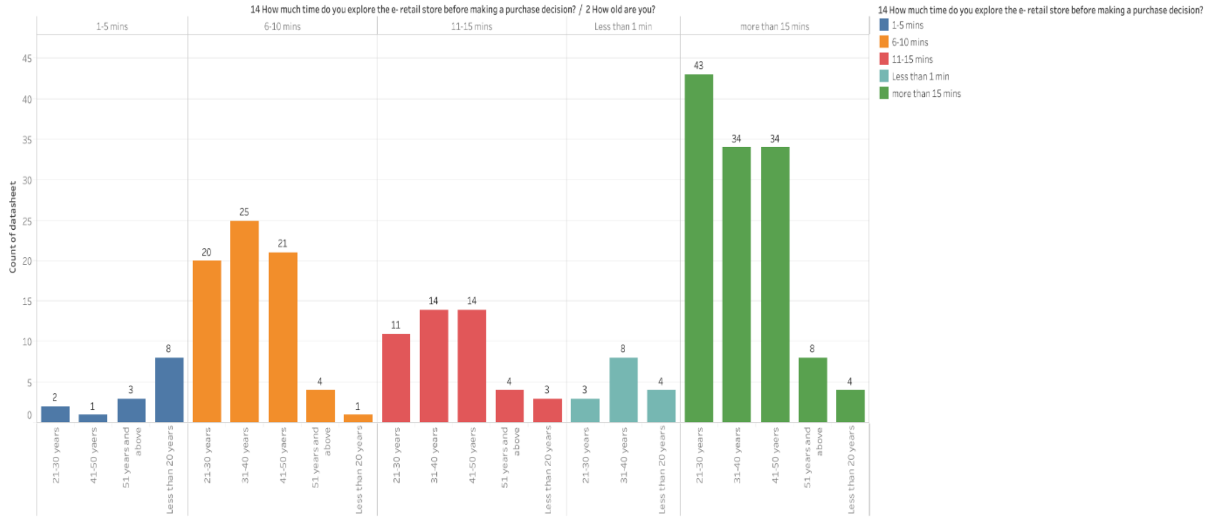
Finding 5: So, among all the customers total 141 use Smartphone to access online shopping and among them most of fro age group 21-40 years, then laptop is more used gadget than tablet and desktop.



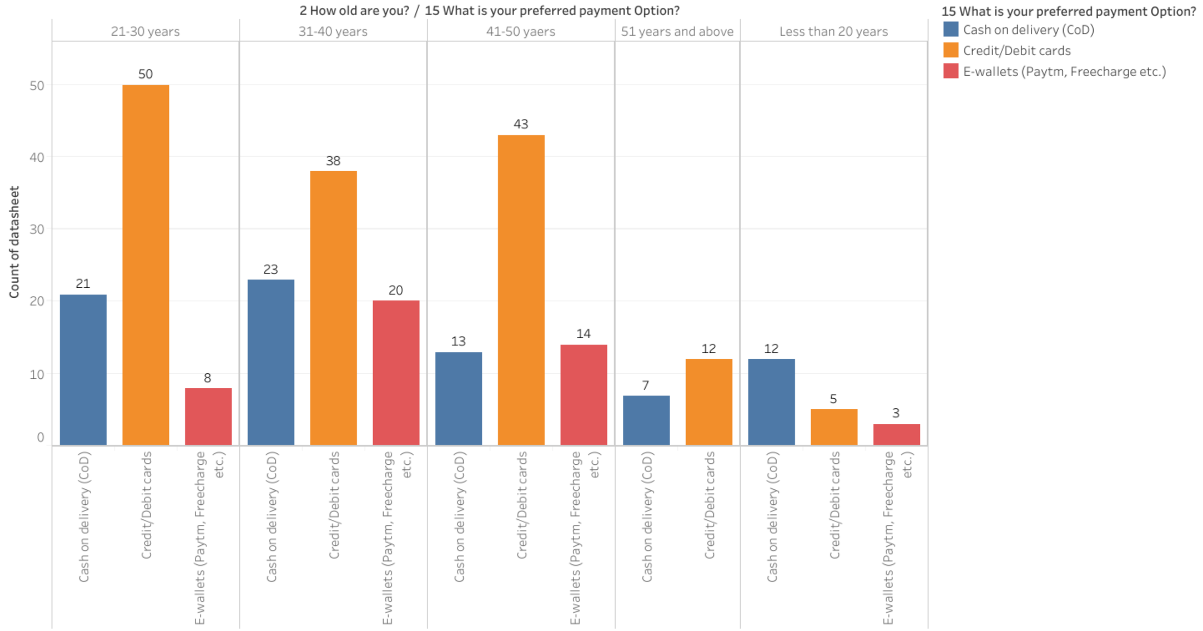
Finding 6: Most of the customers(more than 85%) go to 'Search Engine' to arrive online store for the first time and also through Google Chrome browser.
No customer follow Content marketing channel bu running on Google Chrome.



Finding 7 : More than 40% customers spend more than 15 mins to make a purchase in e-retail store. Among them 21-40 years people are 50% .



Finding 8: Customers between the age 21 and 50 years are more comfortable with Credit/Debit card payment option then Cash on Delivery then E-wallets and less than 20 years of age customers mostly use Cash on Delivery option.



There are few features in the website or application like:

- The content on the website must be easy to read and understand
- Ease of navigation in website
- User friendly Interface of the website
- Convenient Payment methods
- Trust that the online retail store will fulfil 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.)
- Shopping online is convenient and flexible
- Return and replacement policy of the e-retailer is important for purchase decision
- Displaying quality Information on the website improves satisfaction of customers
- User derive satisfaction while shopping on a good quality website or application
- User derive satisfaction while shopping on a good quality website or application
- Provision of complete and relevant product information, Monetary savings.

are the most important features according to the customers. 50% users from our sample dataset (min 135 out of 269) are strongly agreed that according to them the website with high quality of these features are their preferred e-commerce websites.

- Among all the e-commerce companies, customers have shopped from Amazon.co for more times than any other e-commerce companies.
- The Second most popular e-retailer company is Flipkart.co.
- Other popular websites are used for shopping are Paytm.com, Myntra.com, Snapdeal.com.
- From our sample dataset, 82 (more than 30%) customers shop from all 5 online retailers (Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com).
- 79 (around 30%) from our 269 customers would like to recommend Amazon.co only to others and 62 people have chosen both Amazon and Flipkart to recommend others for online shopping, then 30 people will recommend both Amazon and Myntra to others.
- Around 50 % customers from our sample dataset have stated that in case of Paytm and Snapdeal, the delivery period is longer, that’s the reason that these two are not in people’s recommending choices.

Interpretation of the Results

Most of the features were having missing values. Numerical continuous variables were right skewed, so I used log normal transformation. There were some categorical variables which were not having much variance. There were some outliers in the data but due to data shortage I have not removed the outliers. Most of the continuous numerical variables were having positive linear relation with the target variable. From Random Forest Regressor R2 score was 87.9, means 88% of the variance of the dependent variable being studied is explained by the variance of the independent variable.

Higher the R2 score means the model is well fit for the data. However, if R2 score is very high, it might be a case of overfitting. Other metrics Mean Absolute Error, Mean Squared Error and Root Mean Squared Error, with gradient boosting these scores are less then compared to other models. If these errors are less that means the model shows less errors.

Conclusion

Key Findings and Conclusions of the Study

According to Marketing Metrics, the chance of making a sale with a new consumer is between five and 20 percent while the odds of selling to a customer who has already made a purchase from you is between 60 to 70 percent. It's noticeably easier to convince existing customers to buy your products.

- All the features or properties that customers are agreed strongly, should be improved and needed more monitoring so that it can reach to customer's expectation level. Some other tools also could be incorporated to make the experience of customers more joyful and unique, which can encourage customers to spend more time in the website.
- Most of the customers are in the age group between 20 to 40 years, so companies can introduce some loyalty program specially designed for this age group customers.
- Analyse the good thoughts from customers about the website, find out updated information about why customers visit your ecommerce site, what they hope to accomplish there, what they're searching for, and also the barriers they find to ordering from you.
- Develop a loyalty program. Make sure your customers know you appreciate their business. Reward them for repeat orders.

Learning Outcomes of the Study in respect of Data Science

The power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important steps to remove missing value or null value fill it by mean, median or by mode or by 0.
Various algorithms I used in this dataset and to get out best result and save that model. The best algorithm is Decision Tree Classification.

Limitations of this work and Scope for Future Work

We have managed out how to prepare a model that gives users for a novel best approach with take a gander at future lodging value predictions. A few relapse strategies have been investigated furthermore compared, when arriving during a prediction strategy. Straight former imply works bring been utilized within our model, something like that that future value predictions will have a tendency towards all the more sensible values. We concocted an approach with use similarly as considerably information as time permits for our prediction system, by adopting those ideas from claiming gradient boosting.

For further improvement we can also make a better model using Artificial neural network (ANN). It is an attempt to simulate the work of a biological brain. The brain learns and evolves through the experiments that it faces through time to make decisions and predict the result of particular actions. Thus, ANN tries to simulate the brain to learn the pattern in a given data to predict the output of that data whether the expected data was provided in the learning process or not.