

**CUSTOMER RETENTION ANALYIS**

Submitted by :

Faiyaz Ansari

**ACKNOWLEDGMENT**

I want to extend my sincere regards to the below mentioned sources and references who helped me a lot in completion of my Project:

Team FlipRobo

Team DataTrained

scikit-learn official documentation

<https://scikit-learn.org/stable/>

geeksforgeeks

https://www.geeksforgeeks.org/

programiz

[https://www.programiz.com](https://www.programiz.com/)

Machine Learning Mastery

<https://machinelearningmastery.com/>

Medium

[https://www.medium.com](https://www.programiz.com/)

**INTRODUCTION**

* Business Problem Framing

We need to analyse the dataset for the given project , the customer feedback for the online shopping websites like amazon,flipkart,snapdeal, myntra etc. has been taken and customer valuable needs has been taken in to account to solve business problem.

* Conceptual Background of the Domain Problem

[Customer retention](https://www.omniconvert.com/blog/what-is-customer-retention.html) refers to the ability of a business to avoid customer defection for a while. It includes all activities that a business engages in to ensure customer loyalty. An eCommerce business must invest in customer acquisition and retention if it wants to succeed. A successful business is one that continuously works towards building a happy customer base.

* Review of Literature

We are required to model the customer retention with the available independent variables. This model will then be used  
by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the  
strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the  
management to understand the pricing dynamics of a new market.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

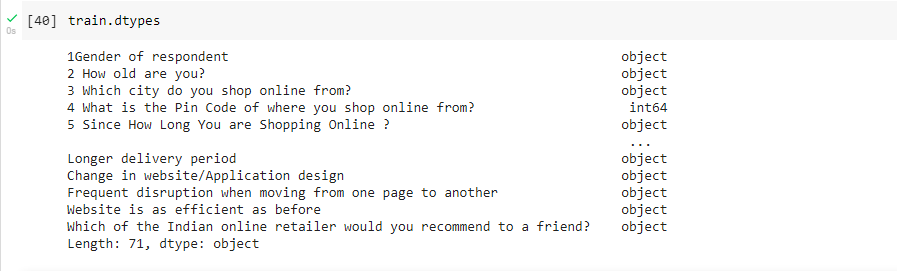
This is a supervised classification problem where we need to predict whether the customer will stay in future or not, so that the company can predict and improve to retain its customers. Thus, we have used classification models like logistic Regression, tree based models like Decision Tree, Ensemble models like Random Forest for the modelling of this task.

The models can be tested on various performance metrices such as confusion matrix, classification report and accuracy score ,f2 score to ensure generalization on future data.

* Data Sources and their formats

The data was provided to us by our client. The dataset contains 269 rows i.e entries and 71

features i.e variables. Below I have attached the datsets with their data types.



* Data Preprocessing Done

The dataset contains no null values . In pre-processing, we have applied several techniques to find what works best. For example, all the outliers were first replaced with the threshold value as per the Inter Quartile Range. Then, we tried removing them as well.

The skewness of the data was removed using various transformation strategies such as log and power transformations of the features.

All the illogical fractional and negative values were either removed or treated with appropriate replacement.

* Data Inputs- Logic- Output Relationships

The data is high-dimensional and can not be visualized directly, but the fact that if classification algos is also performing well on the data indicates that data is almost linearly seperable in higher dimensional space.

We can also used feature importance of Random Forest to have an insight of which features are the most important ones for the classification. Because the data is multi-collinear, we can not use feature weights by linear Regression to estimate a relationship between features and target. But, the pearson correlation gives an idea about how input and output are correlated.

* State the set of assumptions (if any) related to the problem under consideration

We have assumed that any values above and below 1.5 times the IQR is an outlier and will either be treated (replaced) or removed.We have also assumed the features are gaussian distributed and skewness is removed using proper transformation.

* Hardware and Software Requirements and Tools Used
* The size of data is very small, therefore any system running on Windows 7 or higher, Mac or Linux based operating systems with 4 GB of RAM is more than sufficient for the given task. We can use any Python IDE or Jupyter notebooks or Google Colab for modelling.
* Below is the list of tools used for the task:
* sklearn for model building,
* pandas for reading and manipulation of data,
* numpy for numerical operations,
* matplotlib and seaborn for data visualization
* scipy for scientific operations and outlier detection
* joblib for saving the model

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Since the task is classification based on supervised learning, we can use logistic Regression, Tree based classification algorithms, Ensemble models and Nearest Neighbors approach etc.

Since there are a lot of outliers, we should use algorithms which are robust to outliers. Also, the size of data is in lakhs, so we should use algorithms with lower time complexity otherwise training time will be huge. Still, it is not that important factor.

We may also use artificial neural networks but the size of data is not enough, so it has a high risk of over-fitting.

* Testing of Identified Approaches (Algorithms)

logistic Regression

Decision Tree classifiers

K Neighbors classifiers

Support Vector classifiers

Gradient Boosting classifiers

Random Forest classifiers

Ada Boost classifiers

* Visualizations

A lot of plots were made as part of data visualisation. We used libraries like matplotlib, pandas and seaborn for data visualisation.

The key findings are:

There are positive values in all columns.

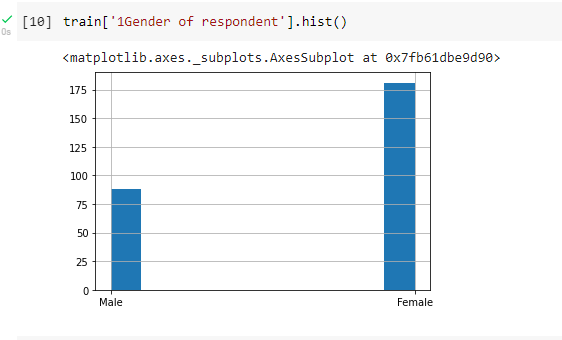
There are exceptionally high values in many columns.

There are few columns which have fractional values.

The data is highly skewed.

Few of the columns are highly correlated.





* **CONCLUSION**
* Key Findings and Conclusions of the Stud

1.females are more than male in the data

2.31-40 age group are more in number and > 51 years are the least

3.delhi is the most shopped city and bulandshahr is the least

201308 i.e greater noida is the most shopped city , 201302 i.e SIKANDRABAD is the least shopped city

4.98 people are shopping online for more than 4 years whereas 16 are shopping for 1-2 years

5.114 people have made an online purchase for less than 10 times in the past 1 year and 6 have made an online purchase for >=42 times in the past 1 year.

#mostly shopping is done via mobile internet

6.smartphones are most use to access the online shopping and tablets are the least use.

7. >5.5 inch phones are the most used

8. windows o.s are the most used i.e 122 and ios/mac are the least used

9.chrome is used by approx all the users

10.customers follow search engine the most to shop online

11.more than 15 mins time customers explore the e- retail store before making a purchase decision by most customers.

12.most payment is done via credit/debit cards.

13.171 customers abandon sometimes shopping after selecting the items.

14.bag item is abandon by 133 customers as they were getting better offer and 54 users abandon the purchase of bag as there were no promocodes.

15.164 customers strongly agreed that the content of websites are understandable..whereas 7 customers strongly disagreed.

16.116 useers strongly agreed that info of similar products is vital.

17.246/269 users agreed that website has ease of navigation

18.234 out of 269 agrees that website has user friendly interface

19.142+86 believes that online retail store will fulfill its part of the transaction at the stipulated time

20.194+42 believes that online ecommerce has empathy towards their customers.

21.185+58 trust that privacy of customers is being guaranteed

22.190 users agrees that Online shopping gives monetary benefit and discounts

23.only 20/269 disagrees that Return and replacement policy of the e-tailer is important for purchase decision.

24.239/269 agrees that user satisfaction cannot exist without trust

25.223/269 agrees they got monetary savings.

26.most people buy products from all the online shopping websites.

27.most no of people who only buys from one website is flipkart i.e 32 followed by amaozon i.e 16

28.flipkart and amazon provides wide variety of product on offer.

29.amazon is the fast loading website speed of website and application.

30.amazon is the most reliable website and application followed by flipkart,myntra

31.amazon has the fastest delivery as per sample followed by flipkart

32.amazon provides privacy of customers’ information more than any other websites.

33.amazon provides the best security of customer financial information as per this sample of 269 customers.

34.amazon is the  most trustworthy website of all.

35.amazon takes longer time to get logged in followed by paytm, myntra , snapdeal and flipkart.

36.myntra has Late declaration of price in terms of promotion and sales period followed by paytm, snapdeal, amazon and flipkart.

37.myntra takes Longer page loading time followed then Paytm.com ,Snapdeal.com ,Flipkart.com ,Amazon

38.paytm takes Longer delivery period and myntra has the fastest delivery.

39.most people will going to recommend amazon to a friend followed by flipkart.

* Learning Outcomes of the Study in respect of Data Science

Data visualization is the utmost important step for any Machine Learning project as it paves the foundation for data cleaning by giving us a detailed insight about the data. It also gives us an idea about which algorithms might work well for the given data. All the insights we received during data cleaning process are listed above. We planned our data cleaning in accordance with those insights.

The main challenge was to clean the data without loss of it. The power of data was utilised in order to formulate proper cleaning strategies. Formulation of proper metrics was crucial for the problem.

Selection of algorithm can be done according to the task in hand. For example, logistic Regression works very well for binary distributed data. Also, there are other factors like Latency, complexity, Interpretability that help us choose the model that can be used.

The model is a little complex, but we don't have a very low latency requirement as per our use case. It is highly interpretable (which is very necessary in our case) as we select features based on Pearson

Correlation.

* Limitations of this work and Scope for Future Work

The provided solution can be made better using more complex models like Artificial Neural Networks. Also, improvements can be made in data gathering and cleaning pipeline, as present data has a lot of discrepencies. We can also construct new features by consulting a domain expert.