Customer Retention Project

Submitted by:

Nikhil Singh Rana

ACKNOWLEDGMENT

The satiation that accompanies the successful completion of the project would be incomplete without the mention of the people who made it possible

I would like to take the opportunity to thank and express my deep sense of gratitude to my data trained academy mentors for providing their valuable guidance at all stages of the study of my data scientist course, their advice and constructive suggestion through which I have gained this much skills that I can complete this project.

I have taken the help of my previous projects which I had done in my training phase with data trained academy and also refered google for some line of codes .

INTRODUCTION

- Business Problem Framing

The first step in not just an ML but any project is to simply define the problem at hand. You first need to understand the situation and the problem which needs to be solved.

So, this problem is related to 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.

- Note : Data Scientists have to apply their analytical skills to give findings and conclusions in detailed data analysis written in jupyter notebook . Only data analysis is required.
  Need not to create machine learning models /but still if anybody comes with it that is welcome.

- Conceptual Background of the Domain Problem

Here the domain problem is related to the Consumer, retail sector. Customer retention refers to the activities and actions companies and organizations take to reduce the number of customer defections. The goal of customer retention programs is to help companies retain as many customers as possible, often through customer loyalty and brand loyalty initiatives.

While most companies traditionally spend more money on customer acquisition because they view it as a quick and effective way of increasing revenue, customer retention often is faster and, on average, costs up to seven times less than customer acquisition. Selling to customers with whom you already have a relationship is often a more effective way of growing revenue because companies don't need to attract, educate, and convert new ones.

Companies that shift their focus to customer retention often find it to be a more efficient process because they are marketing to customers who already have expressed an interest in the products and are engaged with the brand, making it easier to capitalize on their experiences with the company. Infact, retention is a more sustainable business model that is a key to sustainable growth. The proof is in the numbers: according to studies done by Bain & Company, increasing customer retention by 5% can lead to an increase in profits of 25% – 95%, and the likelihood of converting an existing customer into a repeat customer is 60% – 70%, while the probability of converting a new lead is 5% – 20%, at best.

- Review of Literature

This research examined whether specific service and sales skills could improve customer retention rates. A literature review was conducted to examine the following issues: (1) whether customer retention rates could be improved by attempting to resell customers who wished to cancel their accounts or stop services; (2) service quality factors that have been shown

to contribute to customer retention; (3) behaviors and skills that have been linked to service quality and customer retention; and (4) the relationship of specific sales skills to these behaviors.

- ## Motivation for the Problem Undertaken

  The objective behind this project is to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been payed i.e. Non- defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter.

  And the motive of this project is to give the client such a model through which they can easily detect if the person is a defaulter or not by check some conditions, this will help the finance companies to give the loan to the person who will surely return the amount and thus benefit the company.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

1. Prepare the problem

The first step in not just an ML but any project is to simply define the problem at hand. First understanding the situation and the problem which needs to be solved. Once you know the problem well, you then head on to solve it.

Load libraries:

The very first step is to load or import the all the libraries and the packages required to get the results you want. Some very primary and almost necessary packages for Machine Learning are — Numpy, Pandas, MatplotLib, sikit learn etc.

Load Dataset:

Once the libraries are loaded, you need to get the data loaded. Pandas has a very straightforward function to perform this task — pandas.read csv.The read.csv function is not just limited to csv files, but also can read other text based files as well

2. Summarise problem

Okay, so the data is loaded and ready to be actioned upon. But first I need to check how the data looks and what all does it contain. To begin with, you would want to see how many rows and columns does the data have and what all are the data types of each column (which pandas thinks they are).

Descriptive statistics:

Descriptive statistics, as the name suggests, describes the data in terms of its statistics — mean,standard deviation, quantiles etc. The easiest way to get a complete description is by pandas.DataFrame.describe.

Data Visualization :

Data Visualizations are very important as they are the quickest way to know the data and the patterns — if they even exist or not. Your data may have thousands of features and even more instances.

Visualizations using Matplotlib, Seaborn can be used to check the correlations within the features and with the target, scatter plots of data, histograms

and boxplots for checking the spread and skewness and much more. Even pandas has its own built in visualization library — pandas.DataFrame.plot which has bar plot, scatter plot, histograms etc.
So I have performed barplot , histogram and distplot for the visualization purpose

3. Prepare Data
Once you know what your data has and looks like you will have to transform it in order to make it suitable for algorithms to process and work more efficiently in order to give more accurate and precise results. This is essentially Data Pre-Processing which is the most important and the most time consuming stage of any ML project.
Also I have checked the correlation between the columns for better data cleaning purpose
Data cleaning:
Real life data is not arranged and presented to you nicely and in a dataframe with no abnormalities. All this needs to be handled manually which takes a lot of time and coding skills
Pandas has various functions to check for such abnormalities
like pandas.DataFrame.isna to check for values with NaNs etc.
drop irrelevant features using pandas.DataFrame.drop
Outlier Removal:
This is also the main step[to improve the accuracy of the model, therefore I firstly check that how many outliers are present in each column by the method of box plot And then remove the outliers by the threshold method .

4. Evaluate Algorithms :
Once the data is ready,I ll proceed to check the performance of the various classification algorithms .
Split-Out validation Dataset:
Once the model is trained, it needs to be validated as well to see if it really generalized the data or it over/under fitted. The data in hand can be split up beforehand as training set and validation set. This split-out has various

techniques — Train Test Split, Shuffle split etc. You can also run Cross Validation on the entire data set for a more robust validation.
Test options and Evaluation
The models need to evaluated based on a certain set of evaluation metrics which need to defined.
For classification algorithm, some of the common metrics are: Confusion Matrix, F1 Score, AUC/ROC curves etc. Which I have performed after applying the different algorithms.

5. Imporve Accuracy:
After you have the best performing algorithms with you, their parameters and the Hyperparameters can be tuned to give maximum results. Multiple algorithms can be chained as well.
Ensembles :
Multiple Machine Learning algorithms can be combined to make a more robust and optimal model that gives better predictions than the single algorithm. This is known as an ensemble.

- Data Sources and their formats

Dataset is provided by the companies client as our client is in collaboration with the Microfinance Institution who provide financial help and loans to the low income group people, thus they will have the history of previous customers to which they have given loans by arranging the details of the previous customers in the proper format one can apply machine learning to find out the predictions, the dataset include the following columns or the details of the previous customers :

- Data Preprocessing Done

Data usually has a lot of so called abnormalities like missing values, a lot of features with incorrect format, features on different scales etc.

Pandas has various functions to check for such abnormalities, following are some of the functions which I have performed for data cleaning purpose:

- Df.drop, to drop the columns which are not needed.
- Label Encoder, to convert the object columns into the Integer columns.
- Outlier removal, to remove the outliers so that accuracy should be maintained.
- For Visualization, I have used **Dtale** library.

- Hardware and Software Requirements and Tools Used

The hardware and software requirements along with the tools, libraries and packages used are mentioned as follows:

- **Firstly you will need the python in your desktop or laptop**
- **Then import pandas, numpy, csv to read the csv files and work on it.**
- **Now I have imported matplotlib, Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python**
- **Imported Seaborn.sns, Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.**
- **Imported Label Encoder, to encode the object columns into the integer columns.**
- **Imported Dtale to perform visualization, describe, etc.**

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

The approaches I followed, both statistical and analytical, for solving of this problem are as follows:

- Firstly in data cleaning, I drop irrelevant features using pandas.DataFrame.drop, checked if there's any Nan values, removed the outliers, checked the skewness, checked the correlation and then performed the visualization
- I have imported Dtale library for the visualization purposes.

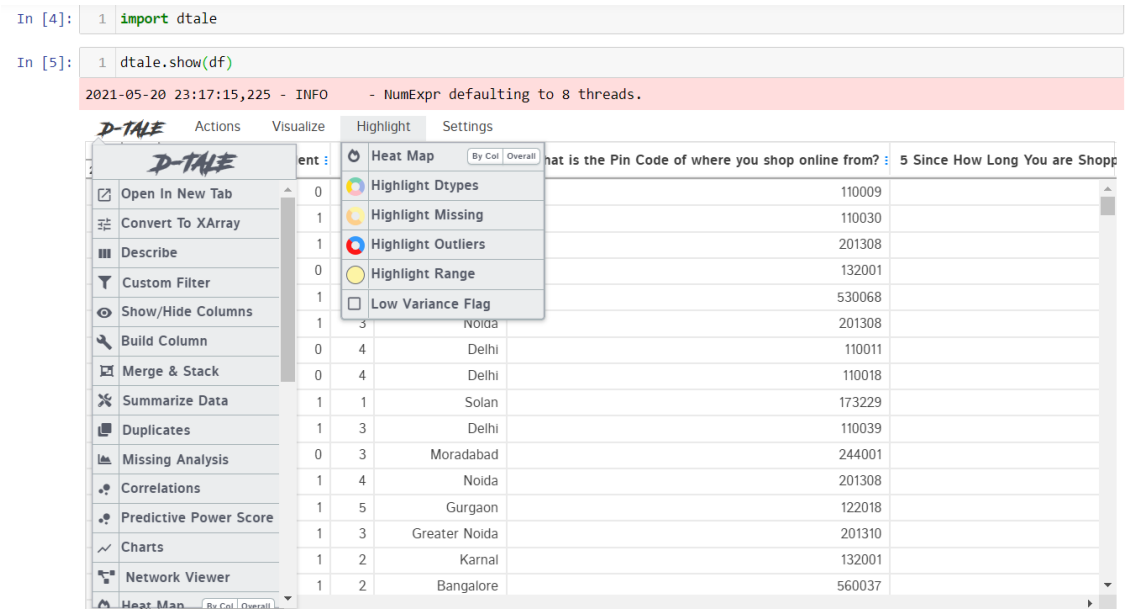
• Visualizations

Data Visualizations are very important as they are the quickest way to know the data and the patterns — if they even exist or not. Your data may have thousands of features and even more instances. It is not possible to analyze the numeric data for all of them .

Visualizations using Matplotlib, Seaborn can be used to check the correlations within the features and with the target, scatter plots of data, histograms and boxplots for checking the spread and skewness and much more.

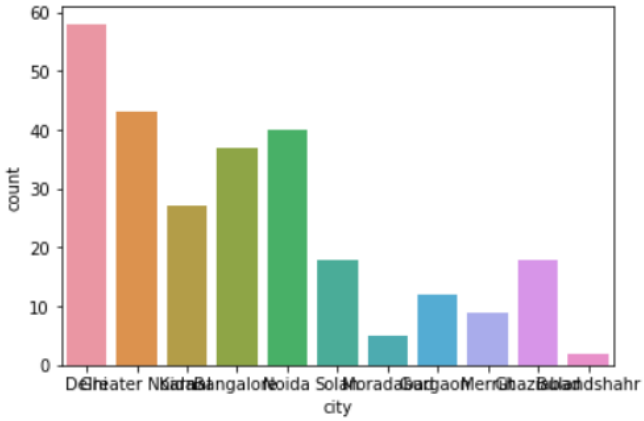
Following are the visualization I have used:

- Dtale Visual Library: With the help of this we can describe the data, find the duplicate entries, find correlation, find the information about each column and try many visual representation.

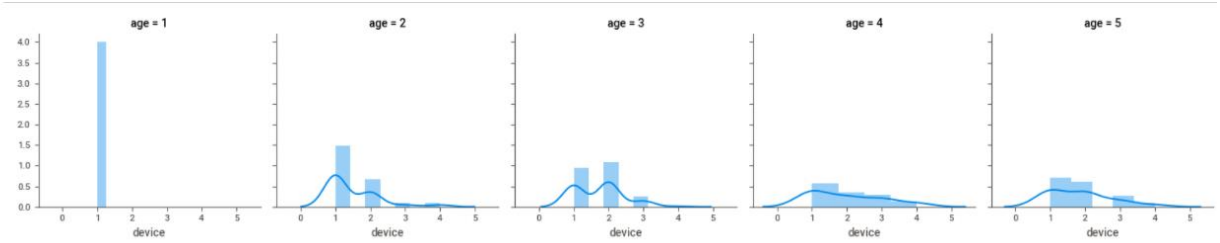


- Count Plot: With the help of it I can see the count of each column For example how many people are from which city as follows:

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x17532194a90>
```



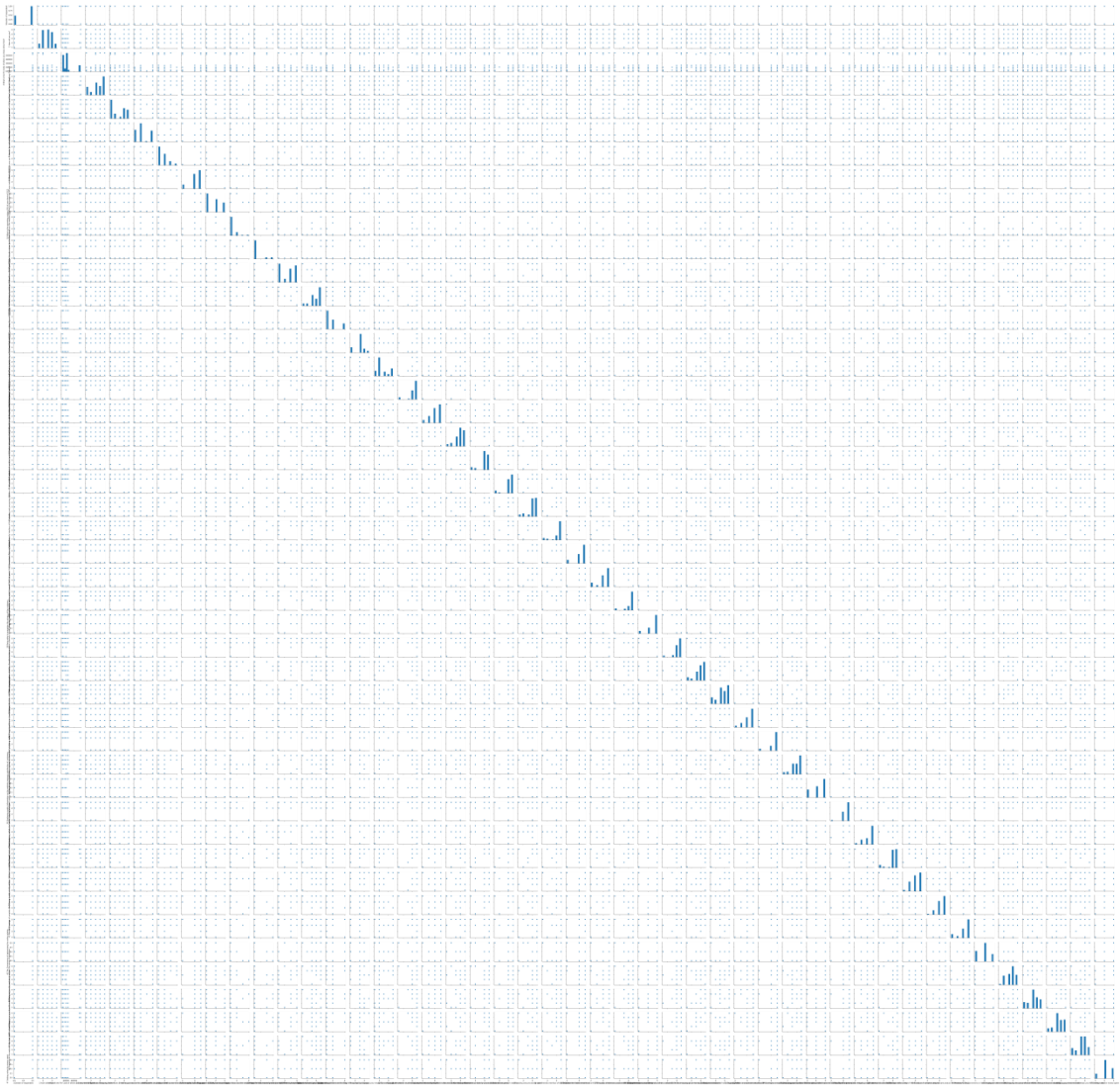
- Here we can see Delhi has more numbers.
- Dist Plot: Seaborn **distplot** lets you show a histogram with a line on it. This can be shown in all kinds of variations. We use seaborn in combination with matplotlib, the **Python** plotting module. ...
  The **distplot**() function combines the matplotlib hist function with the seaborn kdeplot() and rugplot() functions.
  Following we can see relation between 2 different column:



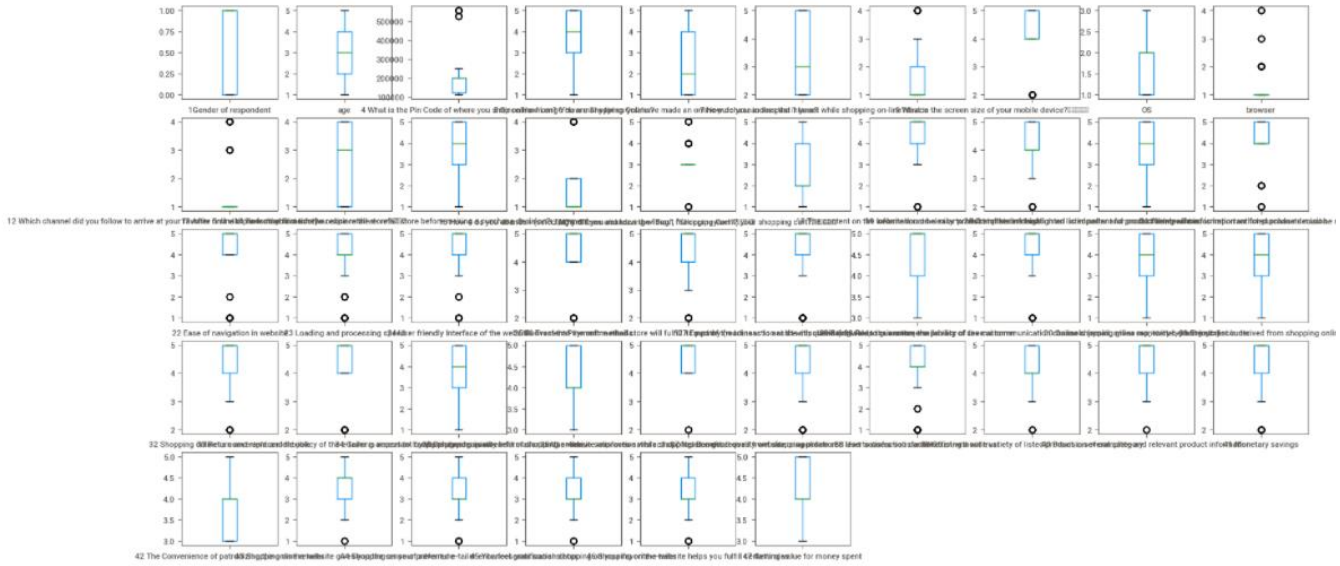
  Here we can see which age group uses which device more. Similarly we can check it for the other 2 columns one by one.

- Correlation: Here we can see the correlation between all the columns with each other

- PairPlot: Here also we can check the relation between different columns as follows:



- Outliers: Here we can check how many outliers are present in each columns, by using BoxPlot as follows:



- Interpretation of the Results

  Summary of what results were interpreted from the visualizations, preprocessing and modeling:

- There are some duplicates entries.
- Most of the Customers are use smartphones.
- Most of the Customers Strongly agree that they have friendly interface of the website.
- Customers who is less than 20 years uses smartphone.
- Customers within age 21-30 uses smartphone and laptops.
- In the Age group which is more than 30 years there are some people who disgree about the information given about the seller is correct .
- There are small no. of outliers in each column.
- "**The Convenience of patronizing the online retailer**" is highly correlated with The **"Shopping on the website gives you the sense of adventure**."
- **"Provision of complete and relevant product information"** is highly correlated with "**Displaying quality Information on the website improves satisfaction of customers**"

CONCLUSION

- Key Findings and Conclusions of the Study

  Key findings are that, the EDA process helps us to understand the problem well through visualizations and data cleaning, this will further help us in model building purpose and gaining the maximum accuracy score.

  Through EDA process we can see that there are some columns which are not needed and there are some outliers to be removed so that the accuracy should be maintained.

  Also there are high correlation between some columns . Therefore we can visualize and check the relationship between that columns.

- Learning Outcomes of the Study in respect of Data Science

  Firstly visualization helps us in various forms as follows:

- **Communicate Findings in Constructive Ways**
- **Understand Connections Between Operations and Results**
- **Interacting With Data**
- **Create New Discussion**

  Secondly, after applying algorithms we can see which algorithm performs better and gives better prediction.

  We can also use ensembling techniques to improve the accuracy of the models build.