Project Report

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

Project Overview:

Ecommerce (electronic commerce) refers to **all online activity that involves the buying and selling of products and services**. In other words, ecommerce is a process for conducting transactions online. When you go to your favorite online retailer to buy a new pair of shoes, you're engaging in ecommerce. The world is moving towards digitalization and that were the idea for Ecommerce comes into the picture.

Sales determine a company's success and profitably. While the world of Ecommerce is highly competitive 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.

In this project we study the factors that trigger a customer to refer a particular shopping website to his friend, which would be an outcome of his satisfaction towards a particular Ecommerce Website.

Problem Statement:

The goal is to create a model that would determine which website depending on a customers preferred choice for shopping would he refer his friend. The task involved the following steps

- EDA
- Data Cleaning
- Train and Test Various Classification Model that determines the website referred.
- Select a Model that produces the highest Accuracy.

The final model selected is supposed to give out the best predictions for the website referred that would be referred by the customer based on his overall satisfaction towards a vendor.

Metrics:

- **Accuracy:** It is a metric that summarizes the performance of a classification model as **the number of correct predictions divided by the total number of predictions**.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

- **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

Methodology

Preprocessing:

- Data Cleaning – Column '3 Which city do you shop online from?' had Noida and Greater Noida which very close by and in the same district which were clubbed into Noida. Column '6 How many times you have made an online purchase in the past 1 year?' had 41 times and above and 42 times and above which were clubbed into 41 times and above. Column's ['4 What is the Pin Code of where you shop online from?','Limited mode of payment on most products (promotion, sales period)','Security of customer financial information','Privacy of customers' information','Speedy order delivery ','46 Shopping on the website helps you fulfill certain roles','44 Shopping on your preferred e-tailer enhances your social status','43 Shopping on the website gives you the sense of adventure','41 Monetary savings','36 User derive satisfaction while shopping on a good quality website or application','29 Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)','25 Convenient Payment methods','24 User friendly Interface of the website','12 Which channel did you follow to arrive at your favorite online store for the first time? ','9 What is the screen size of your mobile device?\t\t\t\t\t\t ','7 How do you access the internet while shopping on-line?','1Gender of respondent'] were dropped.
- Since the dataset was imbalanced, we used SMOTE to balance the dataset.
- Ordinal Encoder was used to encode the categorical variables
- Standardization was performed on Independent Variables.

Model Training:

3 Models were created using cv 4 (LogisticRegression, KNeighborsClassifier and RandomForestClassifier were checked against cross validation scores.

Model Evaluation and Validation:

All the 3 Models returned an Accuracy of 1.0 which is 100%

Cross-validation score of 1.0 which is 100% as below.

```
Report for  LogisticRegression()
The Accuracy score is 1.0
[[17  0  0  0  0  0  0  0]
 [ 0 14  0  0  0  0  0  0]
 [ 0  0 12  0  0  0  0  0]
 [ 0  0  0 17  0  0  0  0]
 [ 0  0  0  0 11  0  0  0]
 [ 0  0  0  0  0 21  0  0]
 [ 0  0  0  0  0  0 19  0]
 [ 0  0  0  0  0  0  0 16]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        17
         1.0       1.00      1.00      1.00        14
         2.0       1.00      1.00      1.00        12
         3.0       1.00      1.00      1.00        17
         4.0       1.00      1.00      1.00        11
         5.0       1.00      1.00      1.00        21
         6.0       1.00      1.00      1.00        19
         7.0       1.00      1.00      1.00        16

    accuracy                           1.00       127
   macro avg       1.00      1.00      1.00       127
weighted avg       1.00      1.00      1.00       127

The Cross-Validation Score is 1.0
Difference in accuracy score and cross-validation score is :-  0.0
```

```
Report for  KNeighborsClassifier()
The Accuracy score is 1.0
[[17  0  0  0  0  0  0  0]
 [ 0 14  0  0  0  0  0  0]
 [ 0  0 12  0  0  0  0  0]
 [ 0  0  0 17  0  0  0  0]
 [ 0  0  0  0 11  0  0  0]
 [ 0  0  0  0  0 21  0  0]
 [ 0  0  0  0  0  0 19  0]
 [ 0  0  0  0  0  0  0 16]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        17
         1.0       1.00      1.00      1.00        14
         2.0       1.00      1.00      1.00        12
         3.0       1.00      1.00      1.00        17
         4.0       1.00      1.00      1.00        11
         5.0       1.00      1.00      1.00        21
         6.0       1.00      1.00      1.00        19
         7.0       1.00      1.00      1.00        16

    accuracy                           1.00       127
   macro avg       1.00      1.00      1.00       127
weighted avg       1.00      1.00      1.00       127

The Cross-Validation Score is 1.0
Difference in accuracy score and cross-validation score is :-  0.0
```

```
Report for  RandomForestClassifier()
The Accuracy score is 1.0
[[17  0  0  0  0  0  0  0]
 [ 0 14  0  0  0  0  0  0]
 [ 0  0 12  0  0  0  0  0]
 [ 0  0  0 17  0  0  0  0]
 [ 0  0  0  0 11  0  0  0]
 [ 0  0  0  0  0 21  0  0]
 [ 0  0  0  0  0  0 19  0]
 [ 0  0  0  0  0  0  0 16]]
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        17
         1.0       1.00      1.00      1.00        14
         2.0       1.00      1.00      1.00        12
         3.0       1.00      1.00      1.00        17
         4.0       1.00      1.00      1.00        11
         5.0       1.00      1.00      1.00        21
         6.0       1.00      1.00      1.00        19
         7.0       1.00      1.00      1.00        16

    accuracy                           1.00       127
   macro avg       1.00      1.00      1.00       127
weighted avg       1.00      1.00      1.00       127

The Cross-Validation Score is 1.0
Difference in accuracy score and cross-validation score is :-  0.0
```

Finalizing the Model:

Since all the model's returned an Accuracy and Cross-validation Score of 100% any model could be picked, I picked LogisticRegression as the Final Model and saved it. No Hyper-Parameter Tuning was required as the best possible score was already achieved.

Reflections:

The process used for the project can be summarized as below:

- The Dataset was cleaned and preprocessed
- A base was created for classification.

- Classifiers were trained.
- Best model was selected and saved.

Finding the correlation was something new for me had to read a lot of articles and found it can be found using Cramer's V method, but the method was not taught very clearly, after some more research found **dython** library by which correlations between categorical and continuous variables can be found. Reducing the features size was another challenge as any number between 3 to 70 for PCA gave me a score of 100% which led to a confusion and was not used.