Problem Statement: For a given set of data by Client ComZ, an ecommerce company, to develop input features data which feed to ML model.

Tools used: Google CoLab, Python

Steps:

1. Data Cleaning
2. Handling Missing Values
3. Feature creation

Pre-processing:

Data Cleaning:

As the data is specific to users, hence dropped the null user data.

Dropped unnecessary columns – Country, City, WebClientID, Browser.

ProductID:

Starting letter of product is in different cases like P and p , hence modified the data to P.

Activity:

Only two options for Activity which are in lower and upper case, hence modified to Upper case.

OS:

Same as above, changed all the values to upper case.

VisitDateTime:

Two kinds of format seen in this variable, one is Unix time format with nano seconds and another datetime with milliseconds.

Logic for Unix time: First 10 characters is Datetime and next 9 characters are nanoseconds, using datetime method, brought the VisitDateTime variable to date format with milliseconds.

Missing Value Imputation:

Created VisitDay variable, format- 'YY-mm-dd' , for simple count of date.

VisitDateTime: Base imputation is Mode(), how?

Taken most repeated Date (e.g., 'YY-mm-dd 00:00:00.000') as per combination, if there are multiple modes then took the latest visited datetime.

1. User, Product
2. User
3. User Segment, Product

Even after imputation, there are few null records, after observation, it is found that users have null Visit datetime at all the instance they logged in, so there is no point in imputing.

Note: Hence filtered the 66 users data for further use of it at the time of feature creation.

Product: Base imputation is Mode, how?

Taken most repeated Product as per combination, if there are multiple modes then took the latest seen product.

1. User, Visitday
2. User
3. User Segment, Visitday

Activity: Base imputation is Mode, how?

Taken most repeated Activity as per combination, if there are multiple modes then took the latest done Activity.

1. User, Product
2. User

Even after imputation, there are few null records, after observation, it is found that users have null Activity at all the instance they logged in; they are 1 time visited users.

Calculated the mode of each kind of users (1 time), found out it is Pageload, hence imputed.

Feature Creation:

No_of_days_Visited_7_Days - Count of days – nunique()

No_Of_Products_Viewed_15_Days – count of products – nunique()

User_Vintage - Date difference of Signup date from '2018-05-28 00:00:00.000'

Most_Viewed_product_15_Days – mode, if more than two, took the latest viewed product

Most_Active_OS - mode, if more than two, took the latest active OS.

Recently_Viewed_Product – Product seen recently, max() of VisitDateTime.

Pageloads_last_7_days – Count of Pageloads – Count()

Clicks_last_7_days – Count of Clicks- Count()

As per requirement, created the features for all the users except those 66 users.

Here, as there is no DateTime at all the instance, so no point in developing features like No of days visited in last 7 days, No of products viewed in last 15 days, Most Viewed Product in last 15 days, Pageloads in last 7 days, Clicks in last 7 days.

However, we can extract the info like, recently viewed product, Most Active OS and User Vintage (Age).