Approach Document for Analytics Vidhya Data Engineering Competition

Objective of this Document is to provide the Insights on the Approach to Generate Features Used, Tech stack Used and Different cleaning and Imputation Strategies Used to Generate the Input Feature Table

Approaches to Clean and Impute missing Data in Visitor DF

Observations and Prerequisites:

1. First validated that for a particular webCLientID, if there is no User ID which is blank, which we might need to impute. Did the validation and found none such users
2. Dropped the duplicates, as there were lot of them.
3. Dropped 'City','Country' which has no usage in Data Engineering process, as to reduce dataframe size

Data Cleaning and Imputations:

VisitDateTime

1. created correctTimeStamp() which will take care of both type of timestamps provided.
2. Imputed values for Datetime using groupby transform as minimum date for the groups
3. Still some NaT value, which were left

Observed Categorical Columns such as **Activity, ProductID, OS, Browser** where data is inconsistent. There are values with both Lower case and Upper-case Values. We need to Convert all of them to either of those cases and make it consistent.

Activity

1. Imputed values for Activity based on UserID groupby transform using **bfill**
2. Remaining values were filled with 'pageload'

ProductID

1. Imputed values for Activity based on UserID groupby transform using **bfill**

Created some binary features which helped in creating input Features eg **SevenDays, FifteenDays, Pageloads_last_7_days, Clicks_last_7_days, pageloads_actvity**

Mostly used groupby agg functions to create the Input Features which took 3-5 minutes to complete the whole Data Engineering task

Created python script for ETL pipeline

Which tools did you use to solve the problem?

Python,Jupyter Notebooks, Numpy, Time and Datetime libraries.

-Rajat Ranjan

-https://rajat5ranjan.github.io/