Approach for Analytics Vidhya Jobathon June-2021

Libraries Used:

Pandas

Datetime

Data Cleaning:

1. Drop Rows with NULL UserID
2. Capitalize Activity Column (to avoid reading 'click' and 'Click' as different entities)
3. Capitalize OS Column (Same OS with different capitalization were present like: 'android' and 'Android')
4. Capitalize ProductID Column (Capitalization error was found in ProductID column)
5. Convert UNIX timestamps to datetime using pd.to_datetime

Data Imputation:

1. Backward fill activity column
2. Merge visitorlogs and userdata dataframes.
3. Sort the final dataframe according to userid

Feature Processing

1. Most Recently Viewed Product:
    a. Remove rows with null VisitDateTime and ProductID
    b. Group data by UserID
    c. Fetch rows with latest date
    d. Remove Duplicates
    e. Merge with submission dataframe
    f. Impute missing values with 'Product101'
2. Most Active OS:
    a. Group dataframe by UserID
    b. Fill submission dataframe with the mode of OS Column
3. Product Views:

    a. Remove rows with null VisitDateTime and ProductID
    b. Group by UserID
    c. Fetch number of unique products
    d. Merge with submission dataframe

4. PageLoads and Clicks:
    a. Remove columns with null date and activity.
    b. Fetch last 7 days data
    c. Group by userid
    d. Get count of pageloads and clicks
    e. Merge with submission dataframe
5. Most Viewed Product:
    a. Remove columns with null date and product id
    b. Filter dataframe for 15 days data
    c. Group by userid
    d. Get count of product ids viewed
    e. Get the most viewed product id and merge with submission dataframe.
6. No. of days visited:
    a. Filter dataframe for 15 days data
    b. Get count of unique days visited
    c. Merge with submission dataframe
7. User Vintage:
    a. Get the difference of the given date(28th may 2018) with the signup date.
    b. Merge with submission dataframe.

Future Scope :

1. Use pyspark to speed up pipeline
2. Data imputation after consulting with the data science team (Random imputation without proper knowledge of data can cause huge affect on ML and DL model’s performance and corrupt their weights).