Approach for the Problem
By: Aman Singh

Data

The data consists of visitors log data which contains the browsing information of all the users and user data which contains the userID and signup date of all the registered users. The two datasets were then merged on userID.

Approach

I have used python pandas to create the input features file. I have used colab notebook for the coding purpose.
First I explored the dataset to know what kind of relationship exists between various columns. It was observed that various columns were having missing values. The DateTime column had different formats of DateTime, and the texts used in various columns were in different cases.

- **Imputing Missing Values:** VisitDateTime column had many missing values. It was observed that every 5 or 6 consecutive values in the VisitDateTime column had the DateTime in the same range so missing values can be filled as the same value just above them. The same was also observed for ProductID and Activity columns. Initially, I filled the VisitDateTime column using “ffill” and ProductID, Activity by taking the mode of the groups made based on VisitDateTime. But finally, I used “ffill” on both ProductID and Activity column as I got better accuracy in the latter case.

- **Different formats of DateTime feature:** The VisitDateTime column contains the DateTime in two formats: The original DateTime format and the other were Unix encoding. The Unix encoding first required to be converted into integer and then it can be easily converted to DateTime format using “Pandas.dataframe.to_datetime”.

- **Values stored in different case:** I have converted all the text values to lowercase so that there will be no discrimination between values belonging to same class but having different cases.

After cleaning the data and filling out the missing values, next task was to obtain the required input features file. I have solved each problem separately by using groupby function intelligently. Output of each of the problem were then combined to form the required input features file.

Kindly see my jupyter notebook where the complete code with approach is explained.