Report

This report is about how and what extra features are created, and merged into a final dataframe.

1. Extract features

multiple_days (0 or 1):

The strategy I used over here is to first assign corresponding day (Day of Month) to each 'uuid'.

Remove all other columns except 'uuid' and 'Day of Month'.

Remove duplicates so that those pairs with same 'uuid' and 'Day of Month' combination are discarded, as we are not interested in those.

Assign **multiple_days** variable TRUE (1) to remaining 'uuid'.

weekday_biz (0 or 1):

Get day of the week (WoD) for each uuid, range of WoD is 0-6. Then keep only those 'uuid' with WoD < 5.

Filter 'uuid' by time for business hours (0900-1700).

Assign **weekday_biz** variable TRUE (1) to those 'uuid'.

Third feature (count):

The purpose of this feature to help predict if a certain 'uuid' is going to buy the ticket or not. The more logins a 'uuid' has done more the probability that a ticket will be bought.

Get count of each 'uuid' by groupby.

2. Merge all the data

Do outer join on 'uuid' to merge dataframe from **multiple_days** and **weekday_biz.**

Then finally merge previous dataframe with original unique set of 'uuid'.

To merge **Third feature,** order the dataframe by 'uuid' and then add additional column.

3. Number of users labeled True

Feature Id	Number of feature labeled True
multiple_days	36035
weekday_biz	90868
third feature	N/A