Project – Ads Recommender System

Data set

- User_data.csv consisting of the user data collected via user activity over the portal for one month with the following columns -
    a. event_time
    b. user_id
    c. event
    d. channel
    e. user_lat
    f. user_long
    g. origin
    h. ad_id
    i. images_count
    j. ad_impressions
    k. ad_views
    l. ad_messages
- Ads_data.csv consisting of the ads specific information with the following columns -
    a. ad_id
    b. category_id
    c. seller_id
    d. creation_time
    e. title
    f. description
    g. price
    h. lat
    i. long
    j. source
    k. enabled
- User_messages.csv consisting of a validation set portraying the user behavior in the following 7 days -
    a. user_id
    b. category_id
    c. ads

Expected Output Format

- Ads_recommendation.csv consisting of the following columns-
    a. User_id
    b. Category_id
    c. Ads_recommended

Dataset Analysis

1. Dataset seems to have been sampled out by location as most of the latitudes and longitudes were very close to each other .
2. Fields like creation_time and event_time needs to be converted into the standard datetime format for better analysis.
3. Fields containing NaN values can be filled by the mean of the column or via interpolation techniques.
4. Columns not containing int64 and float64 datatypes needs to be label encoded in order to be fed to the models. Eg – the field event can be label encoded to the format – {“first_message”: 0, “view”:1} etc.

Approaches

- Recommend the 10 most popular ads to all the users.
- Recommend category-wise most popular ads to all the users.
- Merge the user-data and the ads-data and feed it to a classifier.
- Recommend based on the combinations of personalized and popular choices –
    - Most popular ads in that category
    - Ads recently viewed by the user
    - Ads on which the user has recently messaged
    - Personalized recommendation to the user based on the classifiers prediction.

  A scoring algorithm can be devised to choose the best out of the set so formed.

Further Enhancements

- Use user created features like lat_diff, long_diff defined as follows –
    - Lat_diff = abs(lat – user_lat)
    - Long_diff = abs(long - user_long)
- Use interpolation techniques rather than replacing Nan by 0.