Problem Statement

Premise

Suppose you're an employee at Saavn as a Big Data engineer working closely with the company's ML team. For better user engagement, you're required to build a system that keeps the users updated based on their music preferences. Suppose, a new track of some particular artist has been released. Now, your responsibility would be to push the notification about this song to the appropriate set of audience. For instance, Baadshah's new track "Tareefan" is released. Now, you would probably like to send notifications about this song to the users who prefer to listen to singers like Honey Singh and Raftaar than to users who prefer listening to singers like Jagjit Singh. Pushing a 'rap song' notification to an admirer of classical music is irrelevant. The user may get annoyed at some time and may even uninstall the app.

Solution

To avoid any such repercussions, it is quintessential to push a song notification only to its interested and relevant audience, which in this case will be youngsters interested in rap music. To accommodate this feature in the platform, you need to segregate the users into different cohorts based on some common characteristics and push the notification about any new song to an appropriate user cohort.

User Cohorts are a group of users who share common characteristics. Users of the same gender, age group and locality are typical examples of user cohorts. User cohorts also help to segment the user base into meaningful clusters for analytics, marketing campaigns, targeting ads and so on. Creating a cohort based on basic features such as gender is pretty trivial. But, a majority of listeners do not have a personal account on the platform. Under such circumstances, personalised information such as age, gender, geographical location etc. is unavailable to the developers trying to build a machine learning solution.

However, you can still segment the users based on their activity on the platform. In other words, you can create user cohorts using a 50-day duration of user clickstream data. You will be provided with the following datasets.

Training

You must build a model that clusters users into various groups. The model must aim at clustering users with similar behaviours or listening patterns together. You will be using the following datasets for training your model.

1. Click-Stream Data- Following are the characteristics of this clickstream data:
    - The data set is present in a CSV file format
    - It has a total of four attributes - “User ID”, “Song ID”, “Date” and “Timestamp.”
    - Approx. size 45GB

2. Metadata- Following are the characteristics of the metadata:
    - The data is present in a CSV file format
    - The metadata maps each “Song ID” to an “Artist ID.”
    - A specific "Song ID" can be related to multiple "Artist IDs". (You can have group songs or duets)

Using these datasets, you need to train a model that will build appropriate user cohorts. Once the clusters have been developed, your aim should be to push the most suitable notification matching the users’ preferences in that particular cluster. The groups must be built with the idea to maximise the clickthrough rate for a notification. Click through rate is a measure that indicates the ratio of the number of users who clicked on the pushed notification to the total number of users to whom that notification was pushed. e.g. suppose “Notification A” is pushed to “Cluster 1” that contains 100 users. If out of all these users, a total of 10 users click on this notification, then the clickthrough rate will be 10/100=10%. You should try to maximise the clickthrough rate for each cluster.

In Saavn, the notifications are pushed with the intention of notifying users about their preferred artists. In other words, notification informs users about the updates from their favoured artists.

Each cluster is targeted with information about only one particular artist. Simply put, each group should be associated with only one artist and, just the notifications related to that specific artist must be pushed to the users within that cluster.

Validation

Once you have trained the model, it should be validated by the efficiency for targeting the highest number of “User IDs” with notifications of their preferred artists. In short, you must try to maximise the click-through rate for every cluster.

At this point, you will be making use of the final dataset, which is notification data.

Notification Data-This data is available in two parts- Notification Clicks and Notification Artist.

1. Following are the characteristics of notification clicks

- The data is present in CSV format
- It has a total of three attributes: “Notification ID” for each notification, “User IDs” who clicked on that notification, and “Date.”
- This data contains the list of "User IDs" who had clicked on the notifications that were pushed to them

2. Following are the characteristics of this notification artist:

- The data is present in CSV format
- It has a total of two attributes: “Notification ID” for each notification and “Artist ID” corresponding to the notification.

You will test your model’s accuracy by calculating the overlap between the “User IDs” of a cluster to which you have pushed a certain notification and the “User IDs” who had clicked on that particular notification.

For example, Consider “Cluster-1” has a total of 4 users- A, B, C, and D. Based on the clusters’ underlying properties you deiced that “Notification X”, which contains updates about some artist, can be pushed to these four users. Out of those four users, only two clicked on the notification. So the click-through rate will be 2/4 = 50%. Your task is to form clusters through which the click-through rate, i.e. the updates about each artist, can be maximised. This can be accomplished by building cohorts with similar sentiments and pushing the users in those clusters with artist updates based on their artist inclination.

In case multiple clusters are linked to the same artist then combine both the clusters

- Imagine, if you form three clusters "C1", "C2" and "C3", specifically for the same artist "A1", then combine the three clusters into one cluster.

Note

- For the project evaluation, we will be asking you to report your average click-through rate over a series of five different notifications: 9553, 9660, 9690, 9703, and 9551
- Each cluster will be associated with only one artist but the vice-versa might not be true. So imagine, if a notification corresponding to an artist "A" is sent to three different clusters, you would need to combine the three clusters and report a common CTR.
- In this problem, you have to group users based on their listening patterns. The data that has been provided to you does not contain any explicit information such as age, genre preferences, gender etc.
- The data provides you with information about users’ behaviour history, and you need to build a model that can learn from implicit features and group similar users together. This problem revolves around training models based on inherent behaviours.