Data Science Game Qualification phase: Music recommendation.

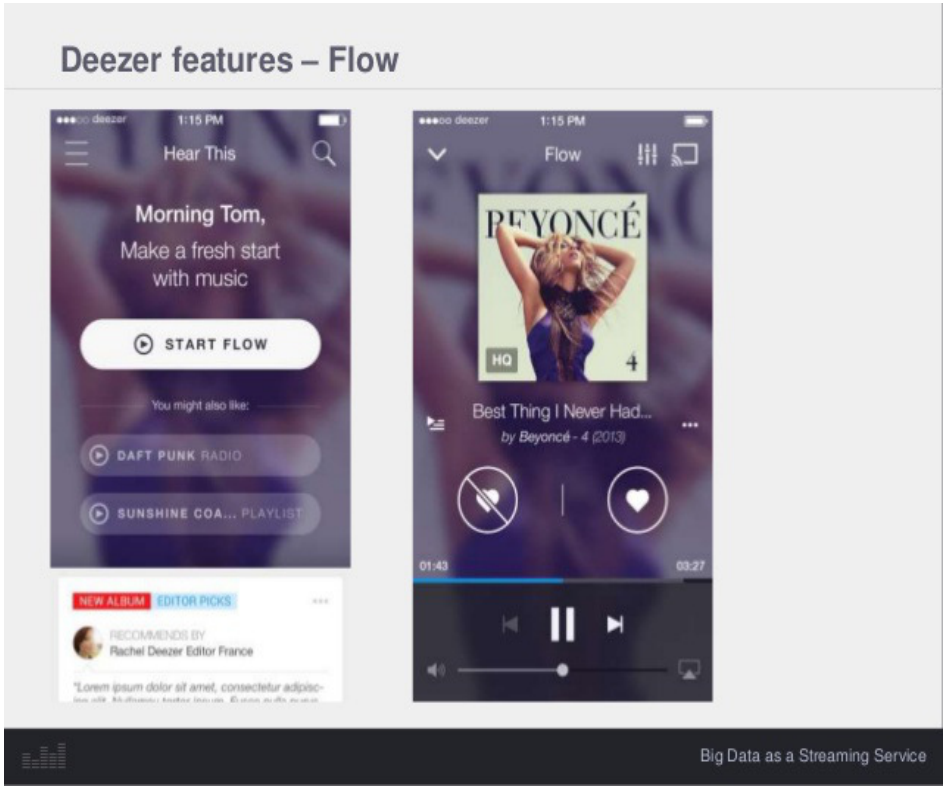
Vikulin Vsevolod Soboleva Daria Popov Nikolay
Shapovalov Nikita

Lomonosov Moscow State University

10 June 2017

Goal

The goal of this challenge is to predict whether the users of the test dataset listened to the first track Flow proposed them or not.



Data

user_id	media_id	f1	f2	...	fn	y
0	211678					1
0	211678					1
0	238286					1
0	615655					0
1	211277					1
...	...	...	...	...	...	...

f1 ... fn – information about user and media.

User-specific features

- user_id – anonymized id of the user
- user_gender – gender of the user
- user_age – age of the user

Media-specific features

- media_id – identifiant of the song listened by the user
- media_duration – duration of the song
- context_type – type of content where the song was listened: playlist, album etc.
- release_date – release date of the song with the format YYYYMMDD
- artist_id – identifiant of the artist of the song
- genre_id – identifiant of the genre of the song

Platform-specific features

- ts_listen – timestamp of the listening in UNIX time
- platform_name – type of os
- platform_family – type of device
- listen_type – if the songs was listened in a flow or not

Artist-specific features

- id – the artist's Deezer id
- name – the artist's name
- radio – true if the artist has a smartradio
- nb_fan – the number of artist's fans

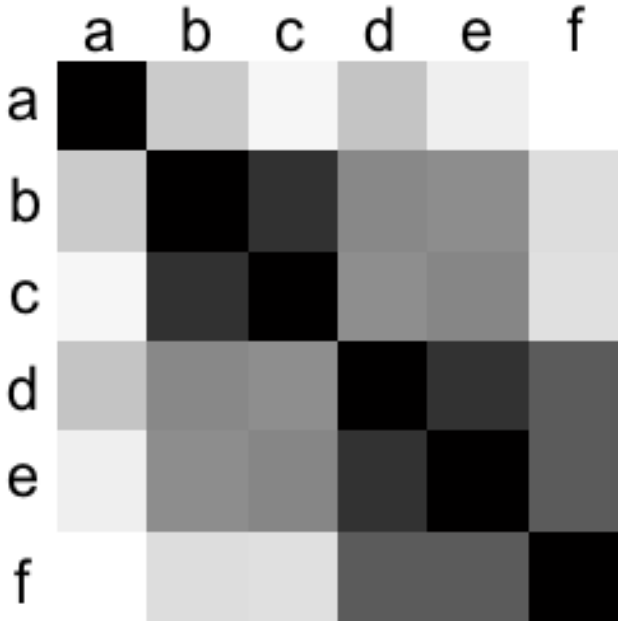
Album-specific features

- id – the Deezer album id
- genre_id – the album's first genre id
- title – the track's fulltitle
- nb_tracks – the number of album's tracks
- rating – the album's rate
- duration – the track's duration in seconds
- fans – the number of album's Fans

Distance matrix features

We create the cosine distance matrix between users.

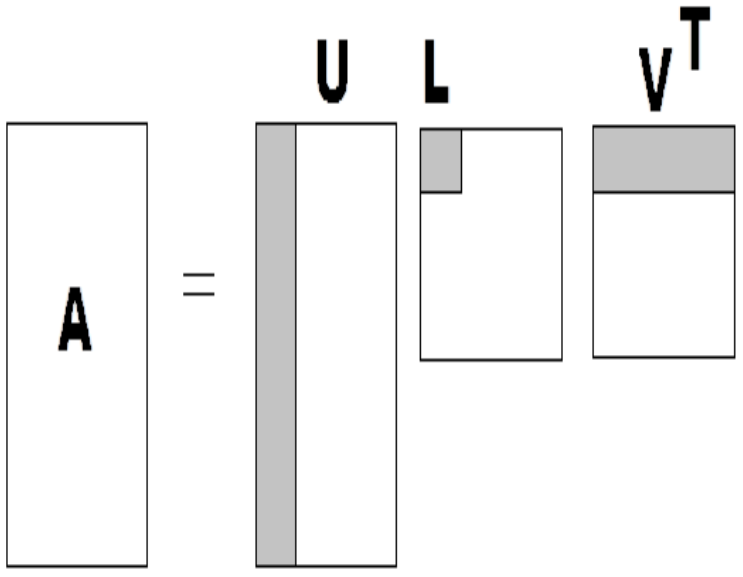
$$distance(user_1, user_2) = 1 - CosSim = 1 - \frac{(user_1, user_2)}{||user_1|| * ||user_2||}$$



Cluster matrix and get the cluster id.

SVD features

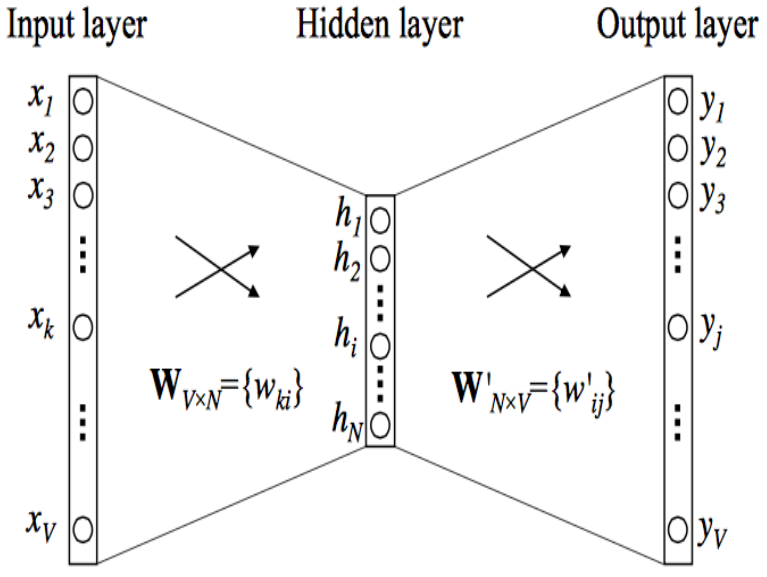
We use SVD for matrix UserMedia.



Cluster users-embedding matrix and get the cluster id or calculate mean of vector for each user.

User2Vec features

We use Word2Vec architecture to predict song for each user.



Get cluster id or calculate mean of user vector for each user.

Solution. Target statistic.

- $K = data.groupby(f).size()$
- $mean\_y = data.groupby(f)['y'].mean()$
- $global\_mean\_y = data['y'].mean()$
- $\frac{mean\_y * K + global\_mean\_y * 10}{K + 10}$

Solution. Real features statistic.

$$data.groupby(f)[real\_f].agg([func])$$

$$func \in min, max, median, etc.$$

Solution. Bug for luck.

$1st$

user_id	media_id	ts_listen	f1	f2	...	fn	y
0	122437314	$1477 * 10^6$					1
0	131340580	$1478 * 10^6$					0

$2nd$

user_id	media_id	ts_listen	f1	f2	...	fn	y
0	131576046	$1479 * 10^6$					1
0	129011934	$1480 * 10^6$					1

$$(1st + 2nd)/2$$

$$\longrightarrow \text{Top1!!!}$$