PROJECT REPORT

Subjects: Big Data Applications: Machine Learning at Scale

Music recommendation system

GVHD: Quách Đình Hoàng

MEMBERS GROUP 2

Nguyễn Anh Đắc
19133020

Nguyễn Thanh Tân Kỷ
19133031

Đào Thị Cẩm Tiên
19133055

Lại Hữu Trác
19133059

CONTENTS

- INTRODUCTION
- DATA
- EDA
- BUILD MODEL
- EVALUATE MODEL AND DEMO APP
- SUMMARY

INTRODUCTION

DATA

- DATASET: Million Song - Recommendation Engines

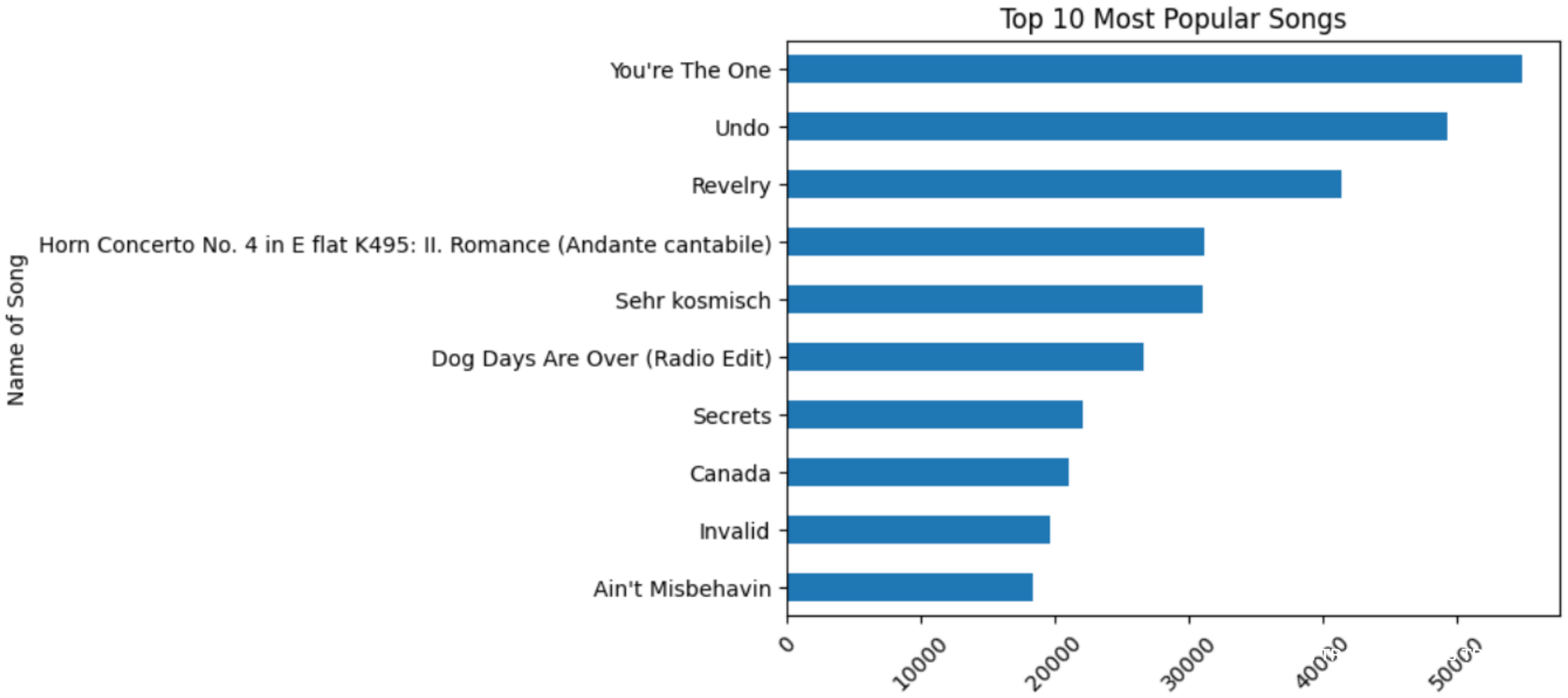
- Song Dataset:
  - song_id
  - title
  - release
  - artist_name
  - year

Size: 1.000.000x5

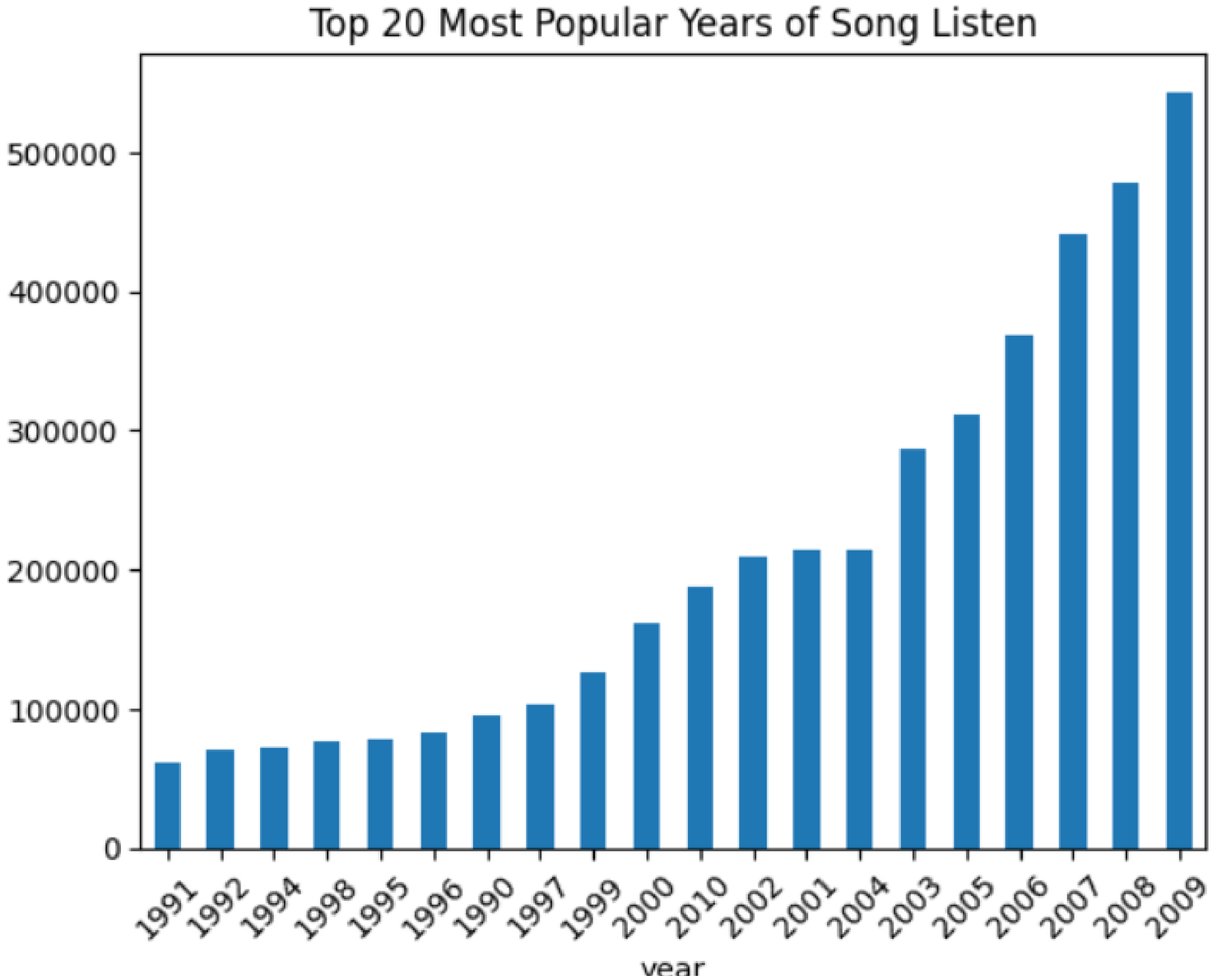
- Users Dataset:
  - user_id
  - song_id
  - listen_count

Size: 2.000.000x3

EDA


Top 10 Most Popular Songs
Name of Song
You're The One
Undo
Revelry
Horn Concerto No. 4 in E flat K495: II. Romance (Andante cantabile)
Sehr kosmisch
Dog Days Are Over (Radio Edit)
Secrets
Canada
Invalid
Ain't Misbehavin
0
10000
20000
30000
40000
50000

EDA


Top 20 Most Popular Years of Song Listen
500000
400000
300000
200000
100000
0
1991
1992
1994
1998
1995
1996
1990
1997
1999
2000
2010
2002
2001
2004
2003
2005
2006
2007
2008
2009
year

BUILD MODEL

- Collaborative filtering
- ALS

$$\mathbf{Y} \approx \begin{bmatrix} \mathbf{x}_1\mathbf{w}_1 & \mathbf{x}_1\mathbf{w}_2 & \dots & \mathbf{x}_1\mathbf{w}_N \\ \mathbf{x}_2\mathbf{w}_1 & \mathbf{x}_2\mathbf{w}_2 & \dots & \mathbf{x}_2\mathbf{w}_N \\ \dots & \dots & \ddots & \dots \\ \mathbf{x}_M\mathbf{w}_1 & \mathbf{x}_M\mathbf{w}_2 & \dots & \mathbf{x}_M\mathbf{w}_N \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \\ \mathbf{x}_M \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_N \end{bmatrix} = \mathbf{XW}$$

$$=> \mathbf{Y} \approx \mathbf{XW}$$

MODEL

Song recommendation system based on the number of listens

- In small set:
  - For training 60%
  - For validation 20%
  - For Test 20%
- ALS value:
  - maxIter
  - ranks
  - regularizations
- Measure: RMSE

MODEL

- Training Model:
  - Input data processing: string-> number
  - Training on training dataset:
    - ranks = [5, 10, 15, 20]
    - regularizations = [0.1, 0.05, 0.01]
  - Evaluate RMSE
  - Choose the best model

MODEL

The best model:

- Rank: 20
- Regularization: 0.1
- RMSE: 6.26

Check on test dataset

- RMSE: 7.28

EVALUATE MODEL

- Recommend Songs for a user

song_id	title	release	artist_name	year	prediction
SOJIPLZ12A6D4F6110	Right Where I Nee...	Greatest Hits	Gary Allan	1999	24.0
SOFJCCE12AB0183F96	Faith	Skunkworks	Bruce Dickinson	0	28.0
SODXXYB12AB0189FA6	Stratus [The Bott...	The Ultimate: Red...	Tommy Bolin	0	28.0
SOGHSMH12A8C137927	Skyway Avenue	We The Kings	We The Kings	2007	33.0
SOFXSLW12A6D4F7BF2	Another Great Divide	The Collection	Split Enz	1981	33.0

- Recommend users for a song

user_id	prediction
5244dd7f3b476a0540153a2720aac03c19c73a20	34.0
0dea223b7ee4abb5fbbde2b47e363e5a4bc8b92f	25.0
99b7b94d6a8f45f674932831d17de4ec90844346	25.0
92527b5bd39847108a05e183cbed2e3633aa40ef	25.0
d94890ce95b60feca22b515e8d7a53b3f1b09f08	24.0

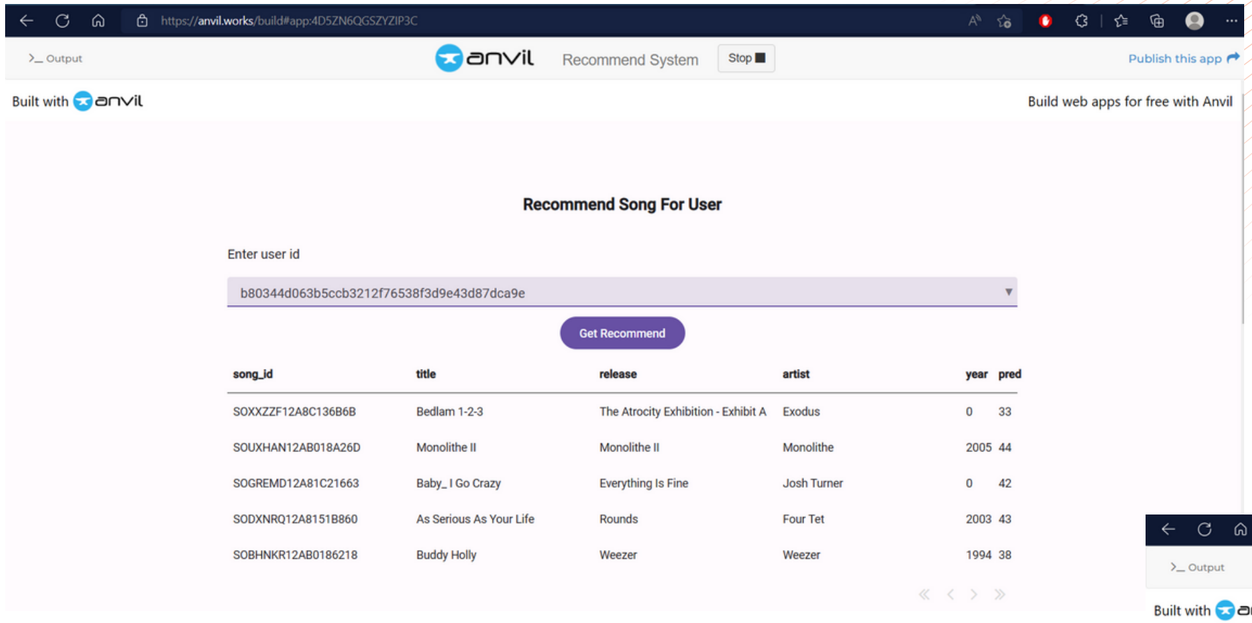
EVALUATE MODEL

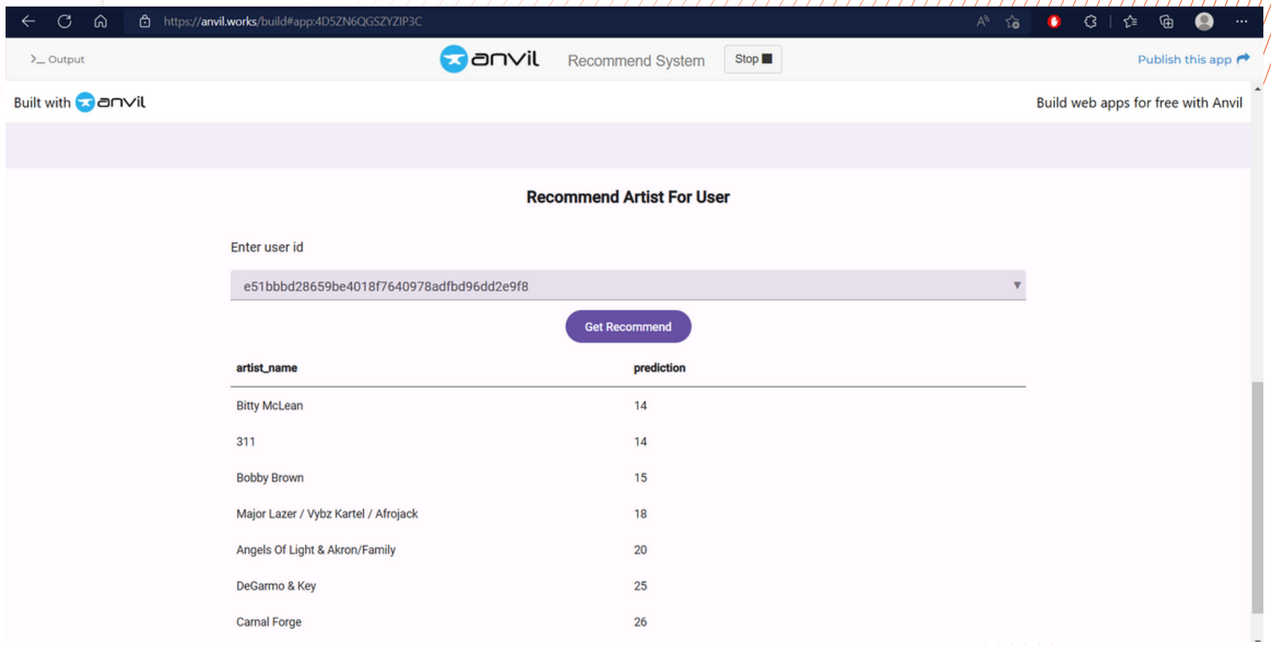
Musician recommendation system for users

- RMSE:
  - Best on Validation: 7.48
  - Test: 10.34
- Recommend musicians for users

artist_name	prediction
moe.	48.0
Black Crowes	53.0
Theatre Of Tragedy	54.0
Savatage	78.0
keller williams	91.0

Web Recommedation





SUMMARY

- SUMMARY:
  - Collaborative filtering
  - ALS

  => The best model.
- The next development direction:
  - Deploy on the website and improve request processing speed, interface

HCMUTE

THANK FOR WATCHING!!!