MUSIC RECOMMENDATION SYSTEM

Project Team: Tulip
Anusha Kalbande: kalbande.a@northeastern.edu
Krina Jitendrabhai Devani: devani.k@northeastern.edu
Neha Joshi: joshi.neh@northeastern.edu
Vidhi Anand Thacker: thacker.v@northeastern.edu

WHAT WE'LL DISCUSS

PRESENTATION HIGHLIGHTS

Introduction
Proposed Solution
Dataset
Methodology
Model performance metrics
Future Improvements
Challenges
Conclusion
References

INTRODUCTION

PROBLEM STATEMENT

The rapid development of mobile devices and the internet has revolutionized the way we access and enjoy music. However, the sheer number of available songs often makes it challenging for individuals to select music that resonates with their tastes.

INTRODUCTION

MOTIVATION

Music service providers face the task of efficiently managing vast song catalogs and helping their customers discover music that aligns with their preferences. This underscores the compelling need for a robust recommendation system.

PROPOSED SOLUTION

We propose to develop a music recommendation system that leverages various Data Mining Techniques and user behaviour data. This system will analyze patterns in user’s listening habits, ratings, and feedback, along with analyzing song features such as genre, artist, and lyrics.

DATASET

The Million Song Dataset is a freely-available collection of audio features and metadata for a million contemporary popular music tracks.

DATASET LINK

hhttp://millionsongdataset.com/tasteprofile/
https://www.kaggle.com/competitions/msdchallenge/data

KEY FEATURES

- Includes information about each song’s popularity, genre, artist, and lyrics.
- Linked to a range of existing resources, making it a valuable tool for research and development.

- 1,019,318 unique users
- 384,546 unique MSD songs
- 48,373,586 user - song - play count triplets

```
Quartile Distributions for play_count:  count    10000.000000
mean            2.522300
std             4.444736
min             1.000000
25%             1.000000
50%             1.000000
75%             2.000000
max           140.000000
Name: play_count, dtype: float64
```

From the above analysis we can see that:

The mean play_count is 2.52 and the standard deviation is 4.44.

25% of the data has play count values of 1 or less and the median of the play_counts is 1.

The play_counts column seems to have a skewed distribution between the range 1-140.

Since the distribution is skewed, we'll apply z-score normalization to the play_count column. This will make sure that the play_count variable now has a mean of 0 and a standard deviation of 1. The distribution will retain its shape, but the scale will be changed.

METHODOLOGY

Our approach will involve learning from users' previous listening histories to provide highly personalized song recommendations. We will implement several techniques:

POPULARITY-BASED MODEL

A baseline model that recommends songs based on their overall popularity.

COLLABORATIVE FILTERING

User-Based (based on user's previous history) or Item-based (songs based on similarity) collaborative filtering can be used to recommend new songs to the users.

METHODOLOGY

SVD (SINGULAR VALUE DECOMPOSITION)

An advanced matrix factorization technique that identifies hidden patterns in user-song interactions.

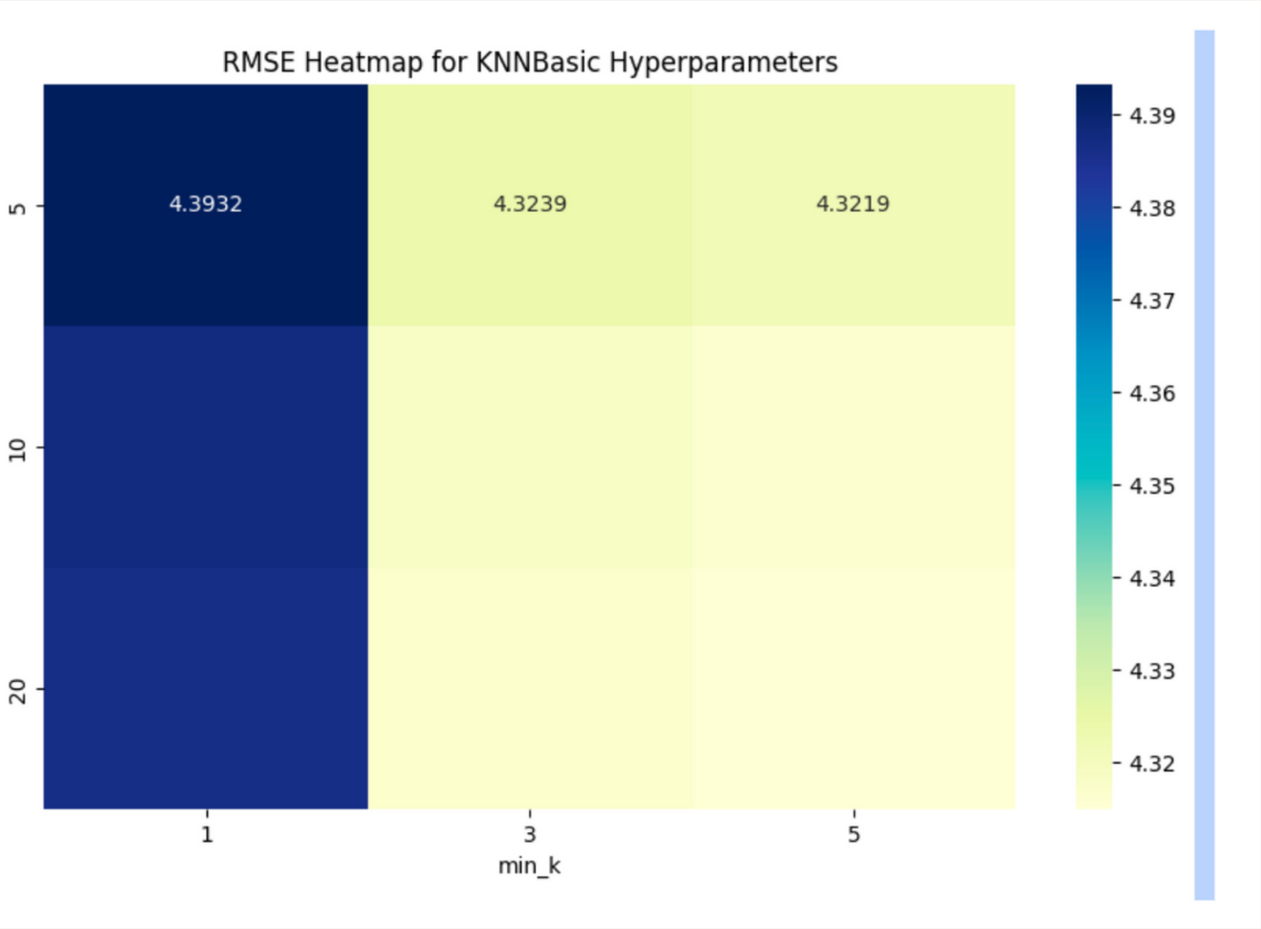
CONTENT-BASED MODEL

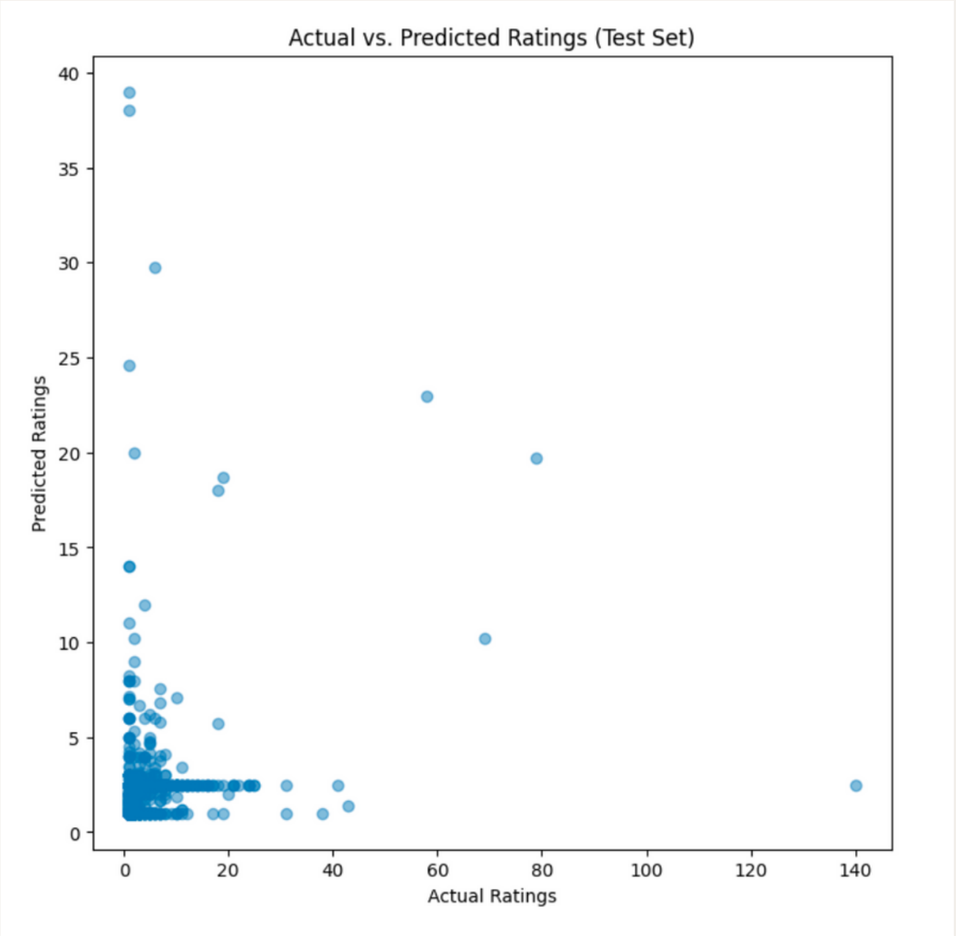
Recommending songs based on their content features and similarity to previously liked songs.

KNN

In a song recommendation system, k-NN recommends songs by finding the ‘k’ most similar songs or user listening patterns.

KNN


RMSE Heatmap for KNNBasic Hyperparameters
4.3932
4.3239
4.3219
5
10
20
1
3
5
min_k
4.39
4.38
4.37
4.36
4.35
4.34
4.33
4.32


Actual vs. Predicted Ratings (Test Set)
Predicted Ratings
Actual Ratings
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100
120
140

COLLABORATIVE FILTERING -STEPS

- CREATING A USER-ITEM MATRIX
- CALCULATE SIMILARITY BETWEEN USERS
- NEIGHBORHOOD SELECTION
- PREDICT THE PREFERENCES OF THE TARGET USER
- GENERATE A LIST OF TOP-N RECOMMENDATIONS FOR THE TARGET USER.
- EVALUATE THE PERFORMANCE OF THE RECOMMENDATION SYSTEM
- FINE-TUNE THE MODEL PARAMETERS
- EVALUATE ON TEST DATA AND GENERATE RECOMMENDATIONS FOR A USER.

COLLABORATIVE FILTERING -RESULTS

- RMSE (COSINE) - TRAIN: 46.748448894575404 VALIDATION: 3.183709094798862 4
- RMSE (PEARSON) - TRAIN: 0.28510219690908095 VALIDATION: 0.9701738159034171
-

TESTING DIFFERENT NEIGHBORHOOD VALUES FOR THE VALIDATION DATA

K VALUE: 5, RMSE-COSINE: 3.183709094798862 4, RMSE-PEARSON: 0.9701738159034171

K VALUE: 10, RMSE-COSINE: 3.183709094798862 4, RMSE-PEARSON: 0.9701738159034171

K VALUE: 15, RMSE-COSINE: 3.183709094798862 4, RMSE-PEARSON: 0.9701738159034171

K VALUE: 20, RMSE-COSINE: 3.183709094798862 4, RMSE-PEARSON: 0.9701738159034171

RMSE (COSINE) ON TEST: 4.337042410553731 RMSE (PEARSON) ON TEST: 1.0934940037702052

COLLABORATIVE FILTERING -RESULTS

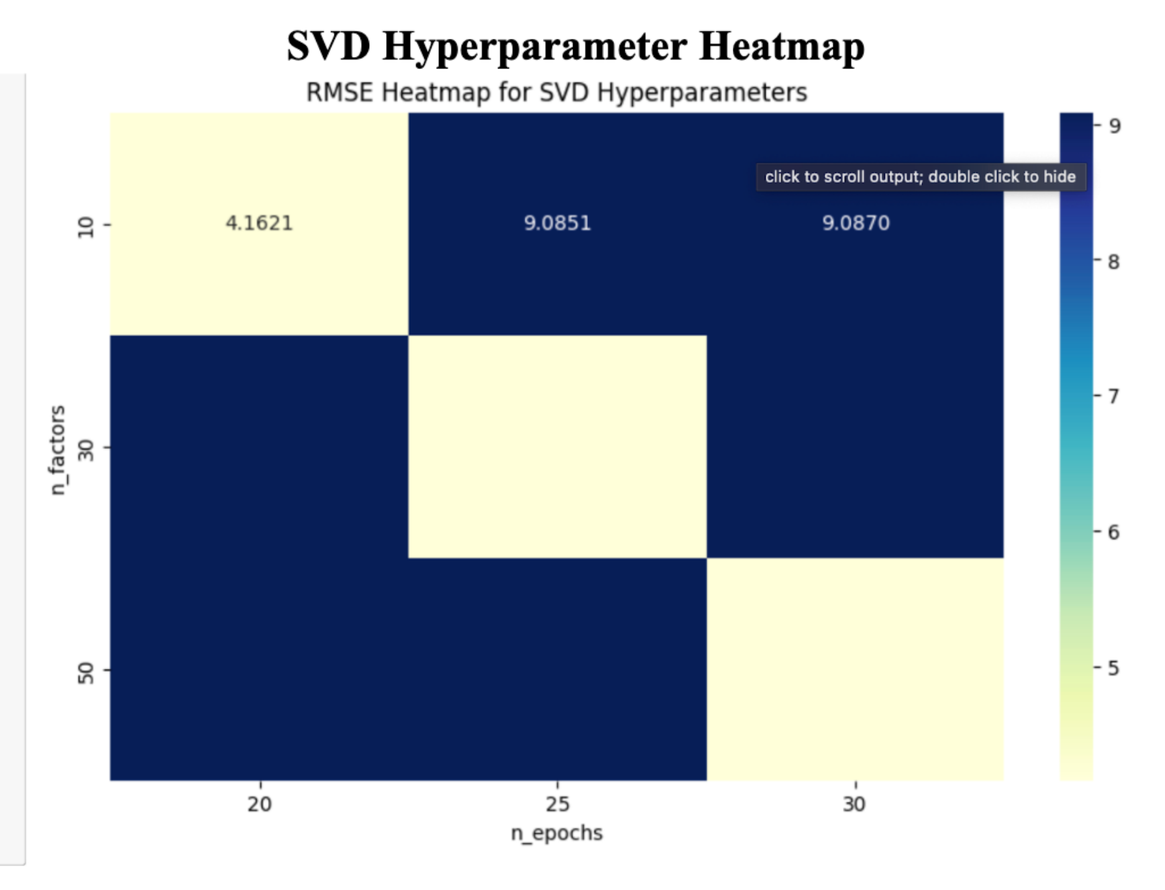
Top Recommendations for User b80344d063b5ccb3212f76538f3d9e43d87dca9e (Cosine):

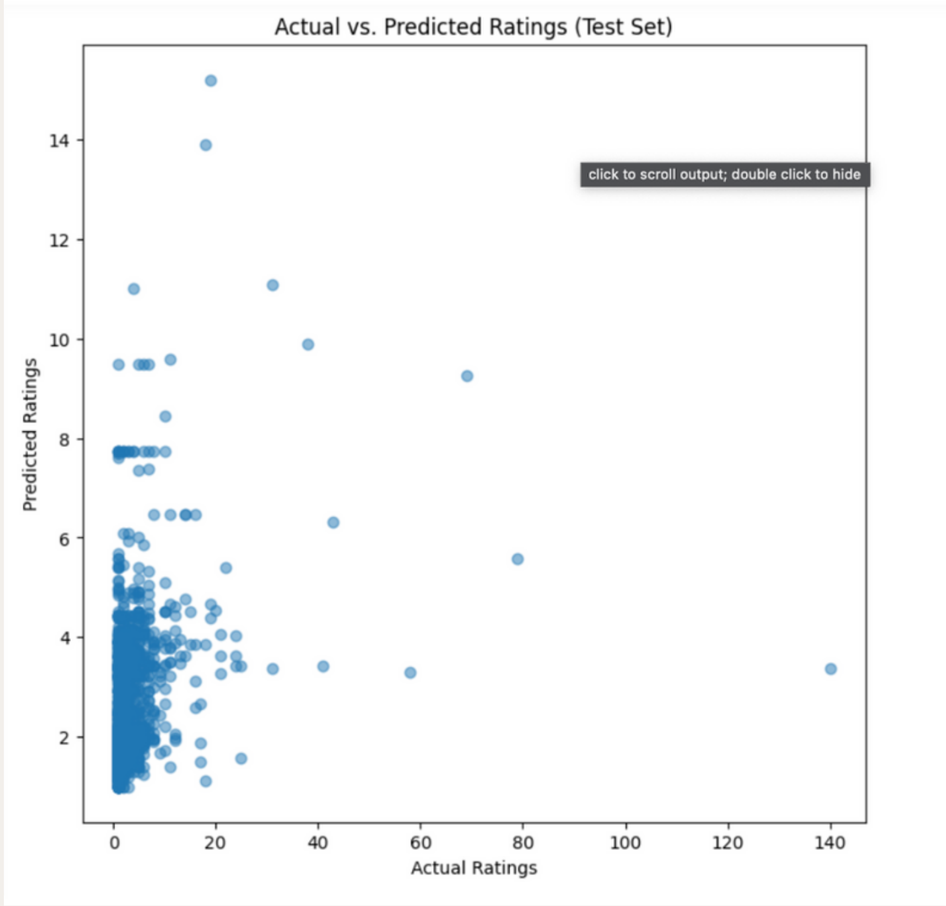
	song	predicted_rating
0	SOAAAGQ12A8C1420C8	33.837002
3158	SOPABZM12A6D4FC668	33.837002
3476	SOQPGDF12AB01858C5	33.837002
5091	SOYTZBN12AB0187A0C	33.837002
3136	SOOXDIJ12A6D4FDE33	33.837002
1693	SOIBCIC12A58A7B55B	33.837002
3009	SOOHWRZ12AC468BA59	33.837002
4539	SOVYNVS12AC3DF64AB	33.837002
4562	SOWBMHX12A8C13851C	33.837002
3545	SOQZYQH12A8AE468E5	33.837002

Top Recommendations for User b80344d063b5ccb3212f76538f3d9e43d87dca9e (Pearson):

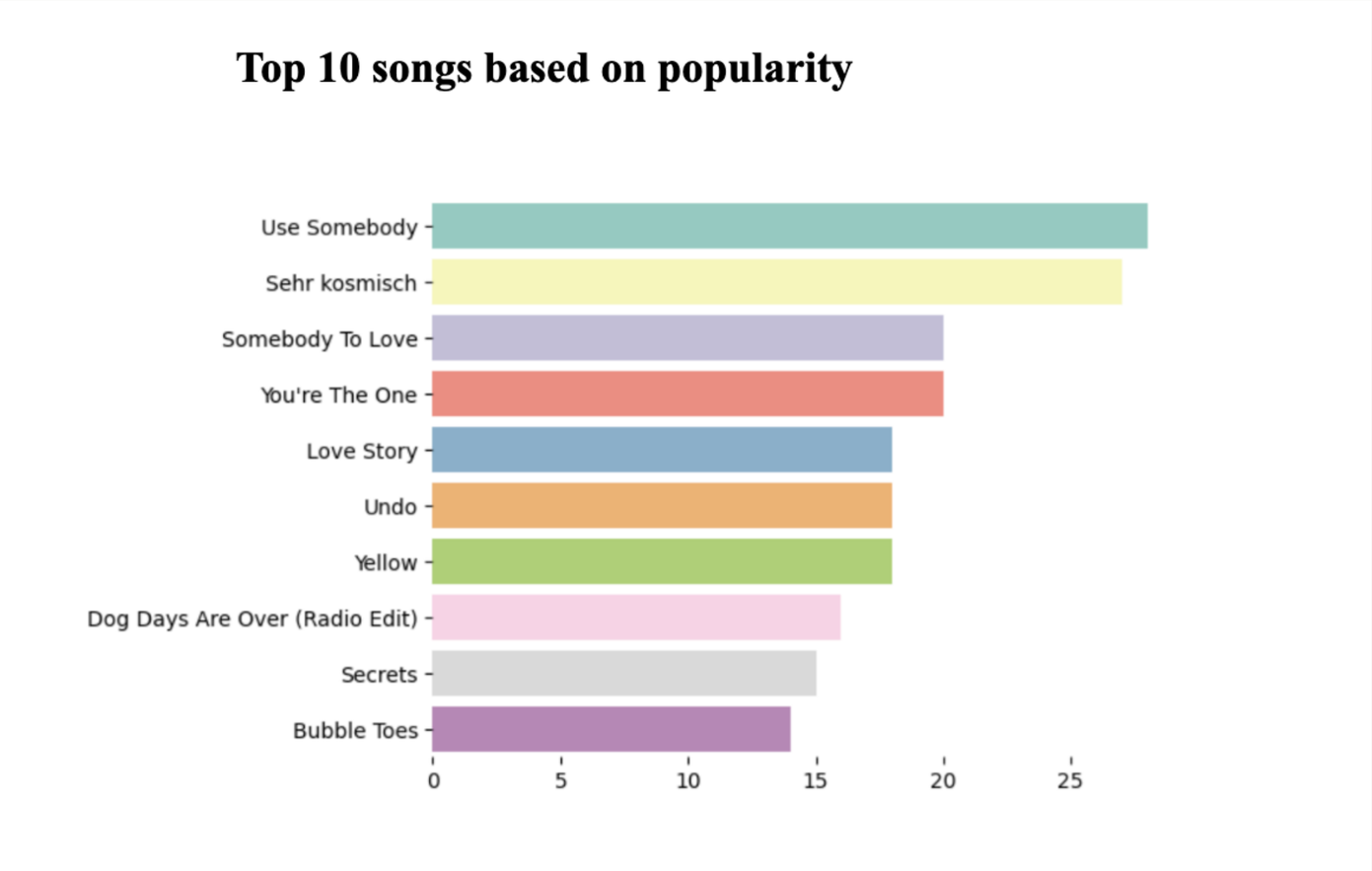
	song	predicted_rating
3545	SOQZYQH12A8AE468E5	6.632379
2552	SOMEIDU12AB0182205	5.282400
1953	SOJIJWG12AAF3B46C0	2.807439
3915	SOSXLTC12AF72A7F54	2.194029
3158	SOPABZM12A6D4FC668	1.907453
494	SOCIHMS12A8C142CC7	1.682456
4467	SOVPBWS12A8C141EF1	1.682456
2565	SOMGIYR12AB0187973	1.503864
4839	SOXKGUD12A58A7C687	0.783873
5027	SOYIZSN12A6701E0BB	0.581311

SVD (SINGULAR VALUE DECOMPOSITION)





POPULARITY-BASED MODEL



CONTENT-BASED MODEL

```
----------------------------------------
Content-Based Recommendation for song at index: The Cove
----------------------------------------

#1: Sehr kosmisch By Harmonia
#2: Stronger By Kanye West
#3: Stronger By Kanye West
#4: Behind The Sea [Live In Chicago] By Panic At The Disco
#5: Learn To Fly By Foo Fighters
#6: The Middle By Jimmy Eat World
#7: Paper Gangsta By Lady GaGa
#8: I?'m A Steady Rollin? Man By Robert Johnson
#9: Champion By Kanye West
#10: Trani By Kings Of Leon
```

FUTURE IMPROVEMENTS

ADDRESSING THE COLD START PROBLEM

REAL-TIME RECOMMENDATIONS

MAKING USE OF IMPLICIT FEEDBACK

CHALLENGES

DATA SPARSITY

Not every user listens to every song, so our data might have a lot of gaps.

SCALABILITY

The more users and songs we have, the harder it is to give recommendations quickly.

COLD START PROBLEM

It’s tough to recommend songs for new users or new songs that don’t have much data yet.

TEMPORAL DYNAMICS

People’s music tastes can change over time, and our system needs to keep up.

THANK YOU