Music Recommendations on Million Songs Dataset

- Sai Prasanna Kumar Kumaru
- Sai Teja Kandukuri
- Neeharika Yeluri

Overview

1. Overview
2. Introduction
3. Datasets
4. Exploratory Data Analysis
5. Algorithms
6. Results
7. Conclusion
8. Challenges

INTRODUCTION

- The Million Song Dataset contains audio features and metadata for one million songs.
- The Taste profile dataset is a subset of the Million Song Dataset that includes user-song-play count triplets, artist and song metadata.
- The project involves data preprocessing, exploratory data analysis, feature engineering, and building and evaluating different recommendation algorithms.
- The project aims to build a personalized music recommendation system that delivers a satisfying music experience to users.

DATASETS

Dataset: Million Song Dataset

MSD Full Dataset:

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

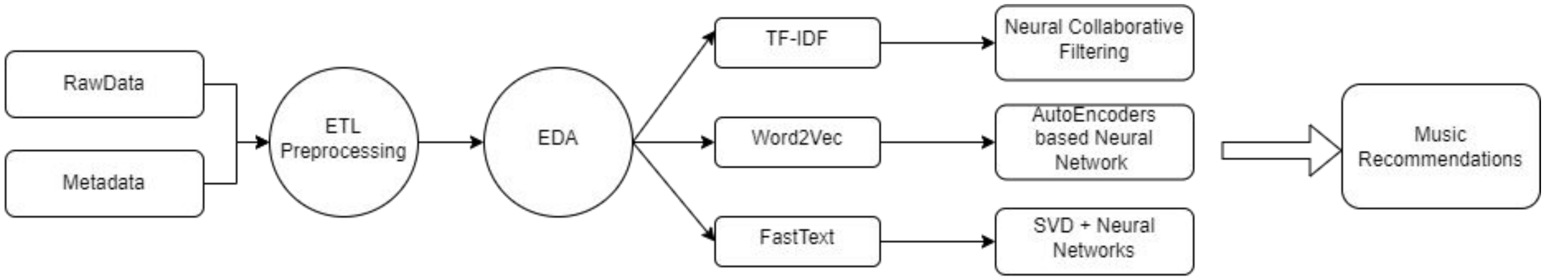
TasteProfile Dataset - Triplets [1] - User ID, Song ID, Play count

Additional Datasets:

- Meta Data [2] -
  - Track ID, title, song_id, Release, artist_id, artist_mbid, artist_name, duration, artist_familiarity, artist_hotttnesss, year, track_7digitalid, shs_perf, shs_work
- Lyrics Data [3] -
  - Track ID, mxm_track_id, word_count

Data Preprocessing

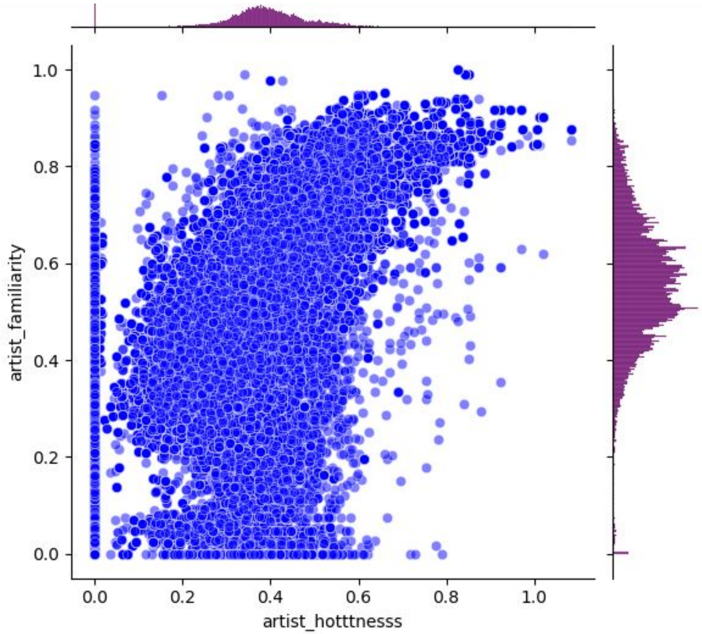
- The data preprocessing stage involved two types of data: interaction data(play count triplets), metadata and lyrics data.
- Outliers in the user song play counts of interaction data were removed using interquantile range.
- The play counts were kept within the range of 1 to 6.
- In the metadata, irrelevant columns were removed and NaN values were also removed.
- Text columns such as release, title, and artist name were preprocessed using NLP techniques.



Exploratory Data Analysis - Song

	title	play_count
0	Use Somebody	22052
1	Undo	16998
2	You're The One	16487
3	Yellow	15262
4	Sehr kosmisch	14624
5	Don't Stop The Music	13492
6	Dog Days Are Over (Radio Edit)	12773
7	Nothin' On You [feat. Bruno Mars] (Album Version)	12740
8	Revelry	12722
9	Bring Me To Life	12144

The most popular song in the taste profile subset is "Use Somebody" with 22052 play counts.

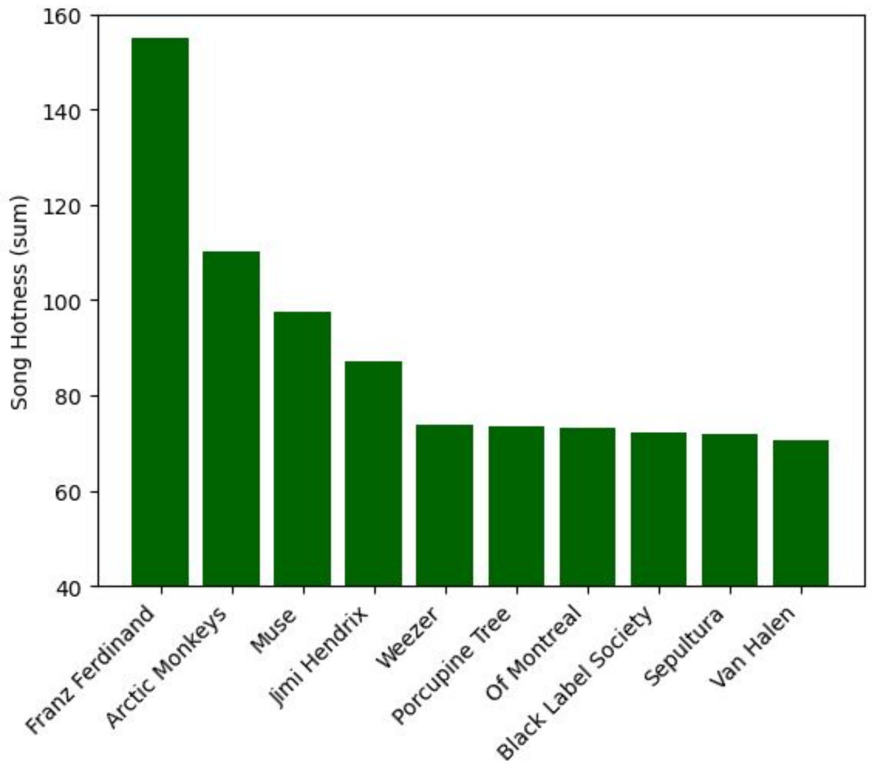


There is a positive correlation between the two variables, indicating that popular artists tend to be more well-known.

EDA – Artists

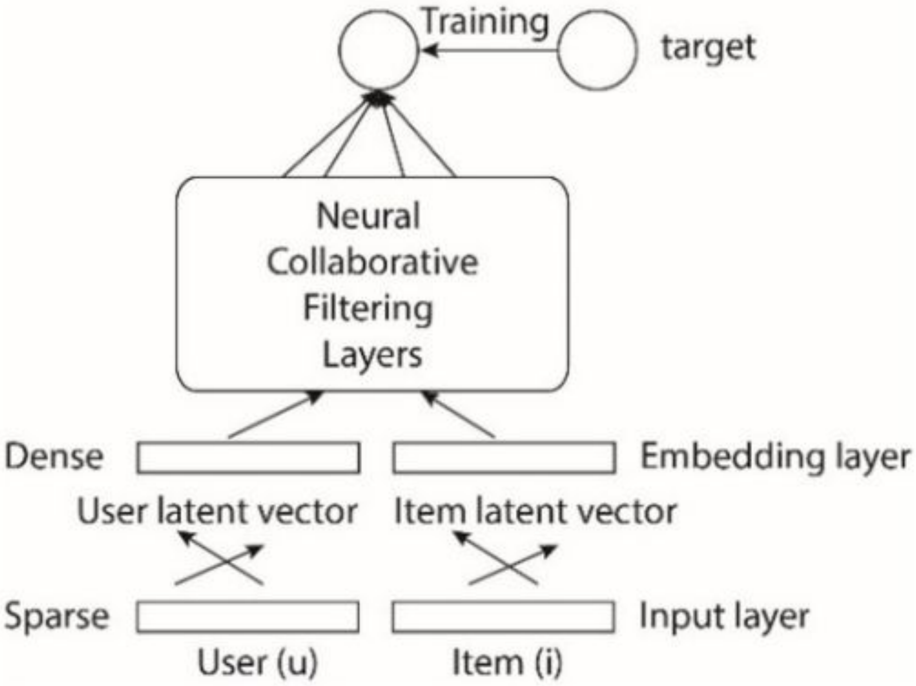
- Top 10 artists based on play count and song hotness.

	artist_name	play_count
0	Kings Of Leon	49797
1	Coldplay	48313
2	Muse	36516
3	Evanescence	30973
4	Eminem	30535
5	Florence + The Machine	28465
6	Jack Johnson	26716
7	The Black Keys	26039
8	Lily Allen	25037
9	Justin Bieber	23671



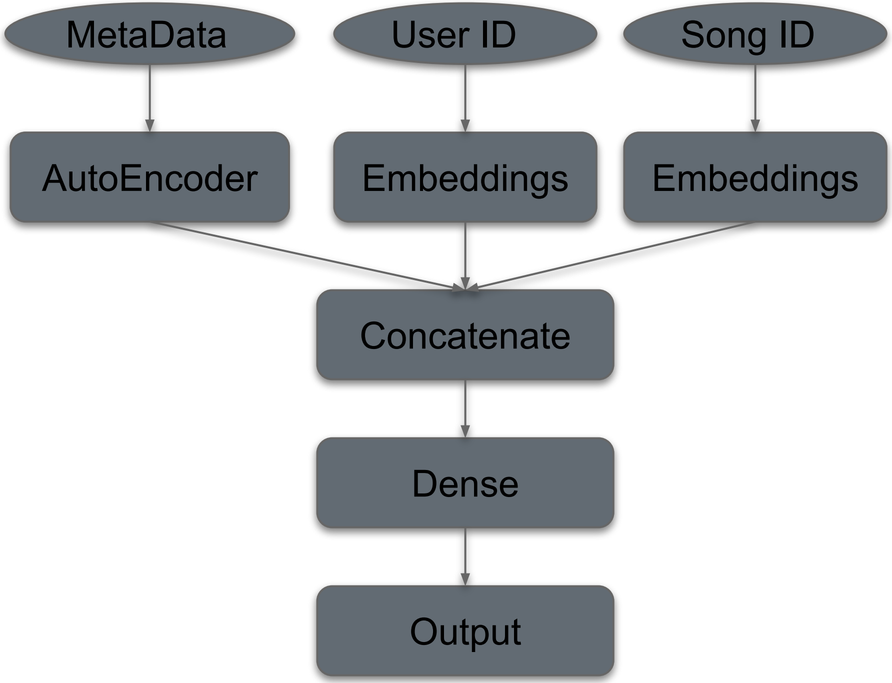
Neural Collaborative Filtering

- A collaborative filtering approach was used to predict song recommendations for a user using play count and song likeness data, as well as metadata.
- The model was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 64, MSE as the loss function and RMSE as the evaluation metric.
- Early stopping was used to prevent overfitting and validate the model's performance.
- song likeness = play count of the song listened by a user / total song play counts of the user



Hybrid Recommender System - Auto Encoder

- A hybrid recommender system using Autoencoders combines both collaborative filtering and content-based filtering techniques.
- Involves encoding user and item features using deep neural networks to generate recommendations, which can handle sparse data more effectively than traditional collaborative filtering techniques.
- The model utilizes both user-song-playcount triplet interaction data and metadata(text and numerical columns).
  - Used Word2Vec model to generate embeddings for text
- An autoencoder is used for the metadata input layer, where dense layers are defined for encoding and decoding metadata information.

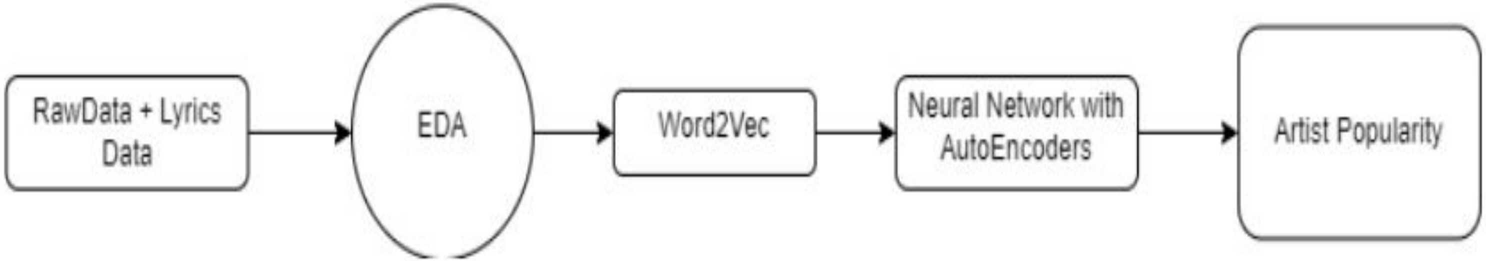


Predicting Artist Popularity

- The metadata is passed through a Word2Vec model to generate embedding vectors, which are later attached to numerical columns to form an array with all metadata information.

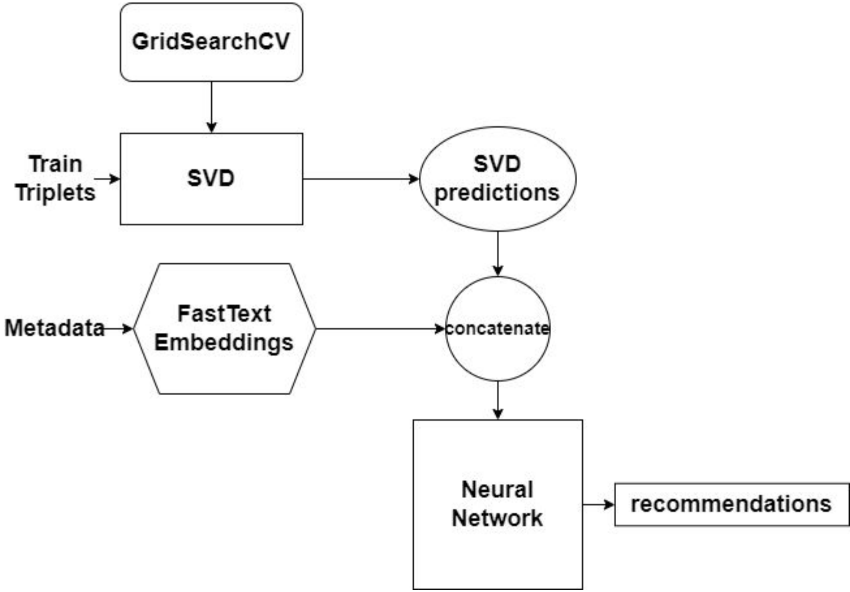
- The encoded layer is merged with the embedding output of interaction data, and the output of the merged layer is passed through several dense layers with dropout regularization.

- The output of the model represents the predicted rating, which can be used to recommend songs to the user based on their preferences and behavior.



SVD and Neural Network

- SVD can learn underlying patterns from data to reconstruct the original data to make predictions for missing targets.
- SVD along with neural network is used to capture complex representations and other factors that SVD couldn't learn.
- SVD is trained from best parameters obtained using GridSearchCV and play_count predictions are generated.
- For the neural network, textual data is encoded using FastText.
- SVD predictions, numerical metadata, textual representations and user item embeddings are trained with the neural network.



Results

TABLE I
RMSE AND MAE FOR ALGORITHMS BELOW ON PLAY COUNT

Algorithm	RMSE	MAE
SVD	1.21	0.87
Basic Neural Network(without metadata)	1.23	0.93
Neural Collaborative Filtering	1.8	1.3
AutoEncoders based Neural Network	1.20	0.86
SVD + Neural Network	**1.14**	**0.82**

TABLE II
RMSE AND MAE FOR ALGORITHMS BELOW ON LIKENESS

Algorithm	RMSE	MAE
SVD	0.029	0.018
Basic Neural Network(without metadata)	0.023	0.10
Neural Collaborative Filtering	0.021	0.107
AutoEncoders based Neural Network	**0.019**	**0.098**
SVD + Neural Network	0.05	0.03

TABLE III
RMSE AND MAE FOR ARTIST POPULARITY MODEL

Algorithm	RMSE	MAE
Auxiliary Output	0.102	0.082
Final Output	**0.015**	**0.0139**

Conclusion

- The SVD + NN approach had the best performance on the recommendation task with RMSE and MAE values of 1.16 and 0.81, respectively, followed by AutoEncoders based Neural Network and Basic Neural Network.
- For likeness prediction task, the AutoEncoders based Neural Network had the best performance with RMSE and MAE values of 0.019 and 0.098, respectively, followed by Neural Collaborative Filtering and Basic Neural Network.
- The neural network model with auxiliary input and output achieved better performance than the model with only a final output layer, and including artist familiarity improved the model's ability to predict artist popularity accurately.
- Incorporating artist familiarity had a positive impact on artist popularity prediction, and this finding could be useful for music industry professionals interested in identifying the factors that contribute to an artist's popularity.

Challenges

- One of the main challenges faced during the development of the music recommendation system was the inclusion of lyrics data due to missing values.
- We had to find a way to address the missing values issue to ensure the model was trained effectively and provided accurate recommendations.
- The dataset used in the project was huge, which made it difficult to implement similarity-based approaches for recommendations.
- Despite the challenges, we explored other options and came up with several effective recommendation algorithms based on the available data.

REFERENCES

1. http://millionsongdataset.com/tasteprofile/
2. http://millionsongdataset.com/pages/getting-dataset/#subset
3. http://millionsongdataset.com/musixmatch/

THANK YOU