Music Recommendation System – Capstone Project

Kedar Kurpad, MIT-PE ADSP MAY24B

Agenda

- Problem
  - How can we improve recommendation accuracy to increase user engagement?
- Approach
  - Develop a system capable of suggesting the top 10 songs for a user based on their preferences
  - Explore advanced models and tuning methods to drive our new recommendation engine
- Solution
  - Deploy a matrix factorization model to improve recommendation accuracy, customer engagement, and user retention (identified drivers of business growth)

Problem Definition

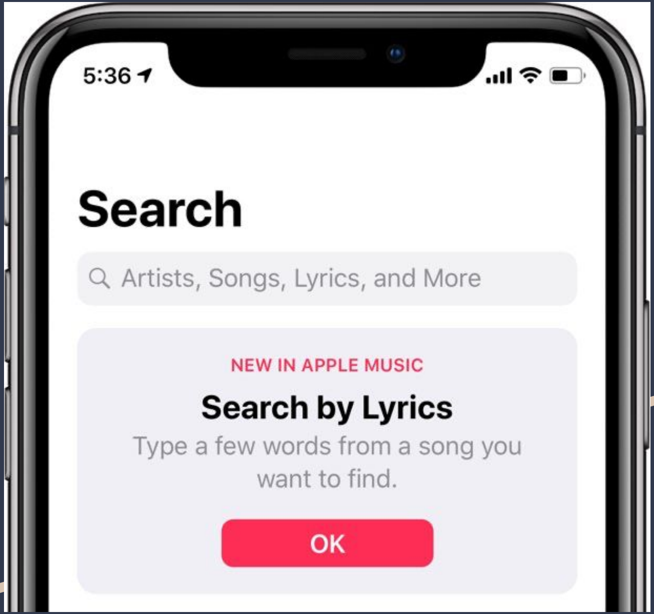
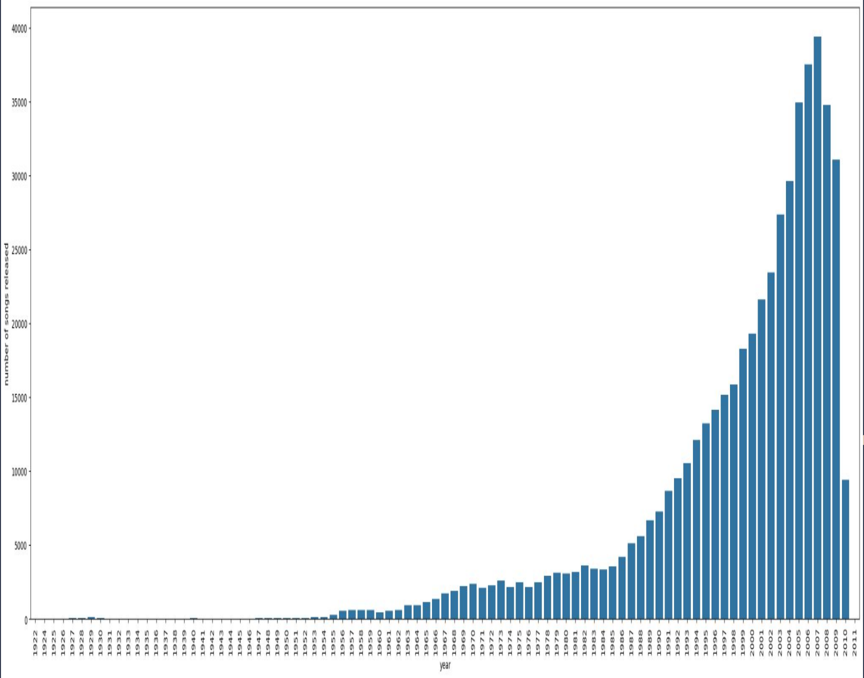


Image Source

- Users struggle to discover new content due to time constraints and an ever-expanding library
- Our company needs an effective recommendation system to improve user engagement and retention metrics

Dataset Notes: The **Million Song Dataset** is a freely-available collection of audio features and metadata for a million contemporary popular music tracks.The core of the dataset is the feature analysis and metadata for one million songs, provided by The Echo Nest. The dataset does not include any audio, only the derived features. Note, however, that sample audio can be fetched from services like 7digital, using code we provide.

Problem to Solve



Barplot visualization of number of songs released each year, using Seaborn.

- How can we improve recommendation accuracy to increase user engagement?
- How can we leverage information from other users' listening habits or from song-specific data to improve our recommendations?

	Unnamed: 0	user_id	song_id	play_count
0	0	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAKIMP12A8C130995	1
1	1	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBBMDR12A8C13253B	2
2	2	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBXHDL12A81C204C0	1
3	3	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBYHAJ12A6701BF1D	1
4	4	b80344d063b5ccb3212f76538f3d9e43d87dca9e	SODACBL12A8C13C273	1

	song_id	title	release	artist_name	year
0	SOQMMHC12AB0180CB8	Silent Night	Monster Ballads X-Mas	Faster Pussy cat	2003
1	SOVFVAK12A8C1350D9	Tanssi vaan	Karkuteillä	Karkkiautomaatti	1995
2	SOGTUKN12AB017F4F1	No One Could Ever	Butter	Hudson Mohawke	2006
3	SOBNYVR12A8C13558C	Si Vos Querés	De Culo	Yerba Brava	2003
4	SOHSBXH12A8C13B0DF	Tangle Of Aspens	Rene Ablaze Presents Winter Sessions	Der Mystic	0

Solution Approach

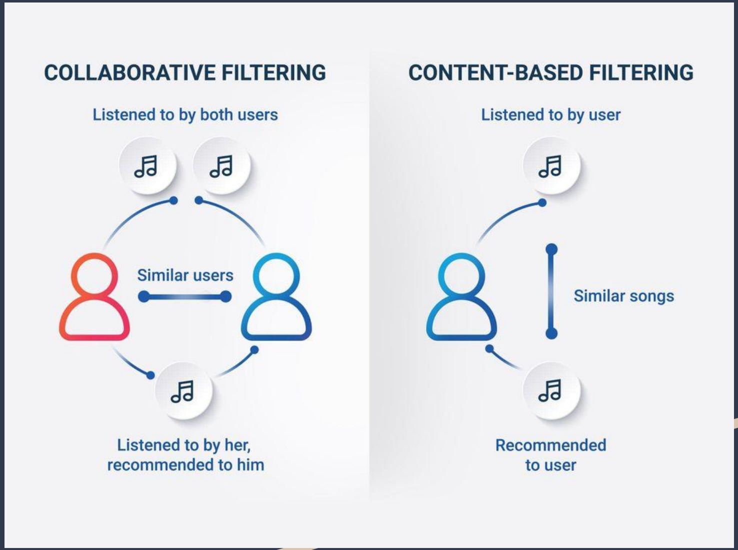
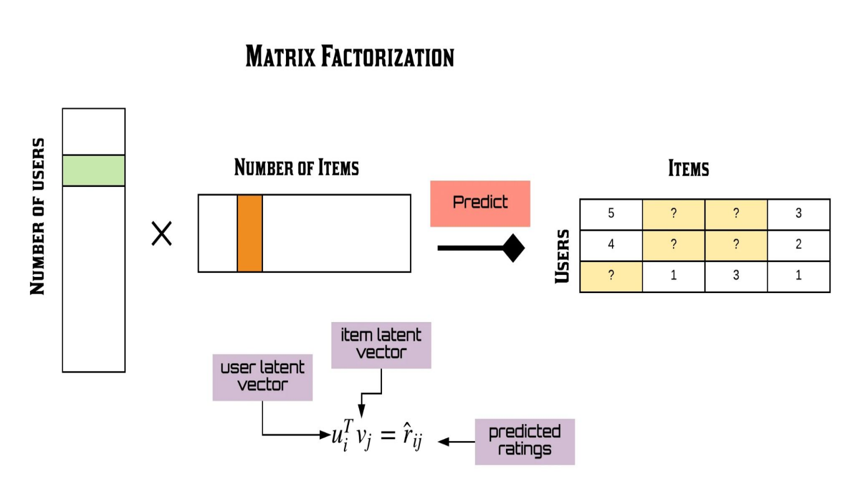


Image Source

- We want to develop a system capable of suggesting the top 10 songs for a user based on their preferences
- We will consider several models including:
  - Popularity-Based Filtering
  - Collaborative Filtering
  - Matrix Factorization
  - Content-Based Filtering
- We will tune our models using GridSearchCV, a method that tests different combinations of settings to find the best possible configuration (a process known as hyperparameter tuning)

Key Findings and Insights



Matrix factorization characterizes items and users using vectors of factors inferred from item rating patterns. High correspondence between item and user factors leads to a recommendation.
Source

- Popularity-Based filtering is simple and highly interpretable, but lacks personalization
- User-User collaborative filtering produces somewhat accurate recommendations, while Item-Item is very error-prone*
- Content-Based filtering offers more diverse recommendations than the Popularity-Based model, but lacks accuracy
- Matrix factorization achieves the best performance, and yields even better results after tuning

During my analysis, my item-item collaborative filtering model produced a suspiciously high RMSE score (in excess of 5), potentially indicating a coding error in my notebook. Further investigation is required on my end to fully understand this error. Nevertheless, matrix factorization produces results that distinguish it as the likely best candidate.

Proposed Model Solution

Model Type	Description	Advantages	Disadvantages	RMSE Score
Popularity-Based	Recommends the most popular songs across all users	Simple to implement	Doesn't consider individual preferences	Not applicable as this method does not predict ratings
Collaborative Filtering (User-User)	Recommends songs based on similar users' preferences	Learns from user behavior, more personalized	Struggles with new users or songs	1.0529 (after optimization)
Matrix Factorization	Factorizes user-item interaction matrix to discover latent factors in preferences	Best performance, captures complex patterns	Computationally expensive, requires tuning	1.019 (after optimization)
Content-Based Filtering	Recommends songs similar to those a user already liked, based on song features	Good for discovering new content within preferred categories	Limited diversity, lacks surprise factor in recommendations	Not applicable as this method focuses on item features rather than predicting ratings

RMSE (Root Mean Square Error) measures prediction accuracy. Lower values indicate better performance. I use it as the primary evaluation metric because it's interpretable, aligns with our rating scale, and emphasizes avoiding large errors in recommendations.

Proposed Business Solution and Potential Benefits

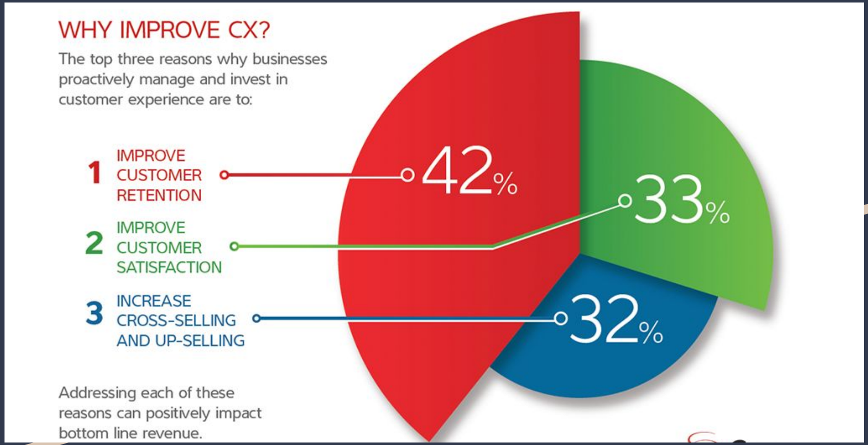


Image Source

- Deploy the matrix factorization model to improve recommendation accuracy
- After post-launch validation of this initial phase, consider hybrid models that incorporate additional methods (such as content-based filtering) to improve recommendation relevance and further improve user engagement
- Increases in user engagement correlate with increases in retention, and therefore drive business growth

Thank you

I welcome any **questions** you may have about my analysis and proposal.