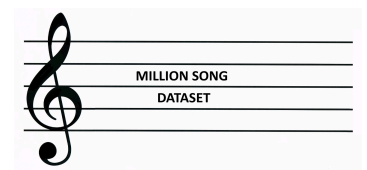# **Music "Tuned"**

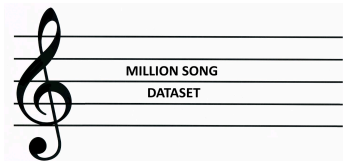## Recommendation System
For the Million Song Database

Prepared by Daniel Holloman
12 January 2026

Contents / Agenda

- Problem Statement / Objective

- Data Description and Overview

- Key Insights from Analysis

- Model Performance Summary / Results

- Exploratory Data Analysis

- Business Recommendations and Next Steps

- Conclusion

- Appendix

Problem Statement / Objective

Spotify purchased EchoNest, a music intelligence platform, in 2014 (Wikimedia, 2025). The Million Song Dataset, which Spotify acquired through this deal, contains song titles, users, metadata, and play counts of songs played by users. Spotify was able to take the raw data from Echo Nest and use it to greatly improve their recommendation systems for their online music streaming platform.

Analyzing the dataset can reveal which songs are most popular. Through careful analysis, the data can be used to predict the most appropriate next song for Spotify users. The dataset contains 2 million data records and 1 million songs. This report uses a cleaned and merged version of that dataset, reducing it down to 400,000 entries in total.

The Million Song Dataset can be analyzed using popularity, user-user, item-item, matrix factorization, and co-clustering models to compare and recommend relevant songs to users. The models predict top songs which the user has not yet played.

Highly relevant song predictions can drive up user engagement and increase subscriptions and ad revenues, increasing profits for the company. Recommended products on ecommerce websites like Amazon can make up 35% of all purchases (Firth, 2025). Other companies such as Netflix and Apple Music also rely on recommendation systems to predict which movies, series, and songs will appeal to their large user base.

Highly precise recommendation systems can help increase sales for many businesses. If customers are happy with recommendations, they tend to stay customers, and they consume more products and services. Recommendation systems are a vehicle for keeping users engaged and consuming music content, leading to higher profits for Spotify.

Over the course of this project, song and user data has been processed in such a way that recommendations can be made to customers. By the end of the analysis, the models can recommend n=10 songs to a user based on their music preferences and their similarity to other songs on the platform.
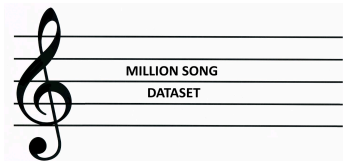
Data Description and Overview

There are 76,353 unique users, 9,999 unique songs, and 3,374 unique artists in our newly filtered database. The data was labeled and the song names and song IDs were encrypted and encoded to numeric features. This resulted in easier analysis, since algorithms work well with numbered data such as integers. Data containing null values were dropped in order to standardize the data and to work with less noise and outliers.

The data set does not include ratings of songs. It only includes playcounts, so we must try to interpret these as an indicator of whether or not the user prefers one song over others. If a user plays a song once, it does not necessarily indicate a strong preference for the song. It could signal the user did not like the song well enough to repeat it.

A threshold must be set. If a user listens to a song more than once, that song is deemed relevant to the analysis of that user. A song play count threshold of 1.5 is set to determine with a greater certainty that a user indicated interest in a song by intentionally playing it more than once.

Most of the songs in the original database were released in the 2000s, with almost 40,000 songs released in 2007. The top played song was "Use Somebody". The max number of songs played in a year by all users was 62,570 in 2007.

Using univariate analysis reveals the count dataframe and song dataframe are composed of integers and objects. There are 17 null values for "title" and 7 null values for "release". The dataframes have a combined size around 100MB.

Using bivariate analysis we can inspect the most played songs in our dataset. "Use Somebody", "Dog Days Are Over", and "Sehr Kosmisch" were all played over 700 times, and are at the very top of this list.

Title	Play Count
Use Somebody	751
Dog Days Are Over	748
Sehr kosmisch	713
Clocks	662
The Scientist	652
Just Dance	623
Secrets	618
Fireflies	609
Creep (Explicit)	606
Yellow	583

User ID	Play Count
75144	621
32542	587
23297	531
10807	512
6480	496
52662	488
61472	482
7834	475
23790	472
49507	468

Key Insights From Analysis

Song releases peaked in the 2000s. The year with the highest number of releases was 2007. The matrix was quite sparse so matrix factorization was used to fill in the gaps by identifying latent factors and identifying patterns and similarities in the data.

Initially, the models were quite slow, and contained many hyperparameters. This resulted in cumbersome models that took more than an hour to run. After reducing the number of hyperparameters, the models became much more efficient and faster. This led to less compute time and reduced cost in regards to electricity required to run the model at market scale.

Root Mean Squared Error calculates the magnitude of error between what is predicted, and what the actual outcome looks like. It is a valuable metric used
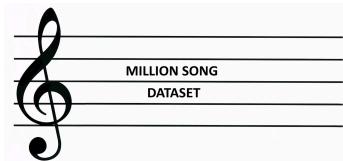
to quantify the errors on a level playing field, and it evens out the observations so that they can be more accurately compared.

GridSearchCV was used to determine several metrics of the test dataset. User-user similarity analysis revealed a Root Mean Square Error of 1.0529. Item-item similarity resulted in an RMSE of 1.0168. SVD RMSE was 1.0193 and Co-clustering RMSE was 1.0760.

Model	RMSE	Precision	Recall	F1-Score
Item-Item	1.0168	0.402	0.556	0.467
User-User	1.0529	0.395	0.621	0.483
SVD	1.0193			
Clustering	1.0760			

Item-item filtering and SVD offered the most accurate predictions. This can be interpreted by the low RMSE scores.

New users with no data (cold start) must be handled with popularity-based recommendations until they grow their playcount above the designated minimum threshold of 90 songs. Only then can they be evaluated with the other models. This ensures we have a large enough sample size to make predictions on songs they may enjoy.

A supplemental Pareto chart (see appendix) identified exponential growth in song releases beginning around the 1980s decade. Then, the song release rate trended downward rapidly in the late 2000s.

Model Performance Summary / Results

The model uses a recommendation algorithm to predict songs that a specified user has never heard before. These recommendations can be based on what the user has already listened to, as well as item-item comparisons, recommending songs related to the songs in their listening history (play counts).

Popularity counts which songs are played the most across all songs in the dataset. The most popular song was "Use Somebody" by Kings of Leon, released in 2008. This type of analysis is useful for cold start problems for new users who have not listened to many songs yet.
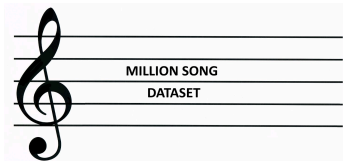
The new user "cold start problem" can be addressed by using popularity based filtering for their first 90 songs. After which, item-item and user-user collaborative filtering can effectively predict relevant songs. However, the data indicates that matrix factorization singular value decomposition was the most effective type of collaborative filtering algorithm.

User-user collaborative filtering was used to determine which users are similar. It stands to reason that users with similar tastes will like many of the same songs. The RMSE was 1.0529, precision was 0.395, recall was 0.621, and F1 was 0.483. The model's hyperparameters were tuned to achieve these results.

Item-item collaborative filtering was tuned and used to compare songs to other songs. This allowed the model to predict new songs for users who have engaged with few songs, but with greater than 90 plays. When Pearson Similarity was employed, The RMSE was 1.0168, precision was 0.402, and recall was 0.556.

When the Matrix factorization (SVD) model was tuned, the RMSE was 1.019, with best parameters coming from running 10 epochs. This is effective for least squares fitting of data. This was the most successful model attempted in this report, in regards to RMSE.

The tuned Co-clustering model returned an RMSE of 1.0760, using 10 epochs. The model performance can be improved over time by increasing the size of the dataset and by continuously re-tuning the models, experimenting with hyperparameters and numbers of epochs. It can also be improved by continuously running the models over time.

Exploratory Data Analysis

There are only about 10K unique songs in the test database. There are approximately 76K unique users and about 3K unique artists in the database. Since the number of unique artists counts is less than unique songs, the data implies that many artists appear in the database multiple times for different
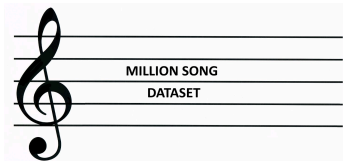
songs they have released. The most interacted song was "Use Somebody" by Kings of Leon with 751 plays. Max user interactions was from user 75144 with a total of 621 songs played.

The count of song releases by year started small in the 1920s, when recorded music was uncommon. Fast forward to 1979, the advent of the Sony Walkman and similar music playing devices spurred growth of song releases as well as sales of albums on cassette and eventually on compact disc.
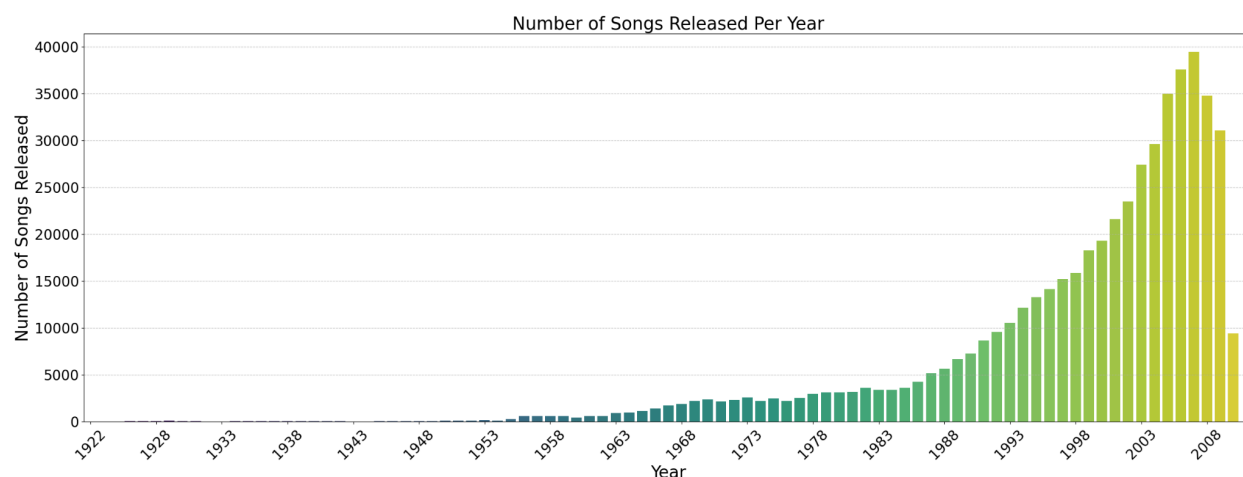

In 1981, MTV took off, raising awareness of musicians as bands scrambled to make music videos, spurring growth of audio recordings. Beginning in the early to mid 90s digital music began to take over. And much of all physical media was digitized across the world. Then around 2001, the first iPod was released, and the online music business exploded. As millions of legitimate purchases were loaded onto iPods, so too did bootleg copies of albums.

In 2005, the advent of Youtube meant that artists could self-publish their musical works and get more exposure than ever before. Then in early 2007, the iPhone was released, putting a personal recording studio in the pockets of many musicians around the world.

The music business has flourished since then, in part because of music platforms like Bandcamp and Soundcloud in 2008. Although creative expression through music was thriving, many consumers and musicians criticized big business for stealing profits from lesser known musicians around the world. All

of these factors contribute to the exponential growth of music releases beginning in the late 80s to early 90s.
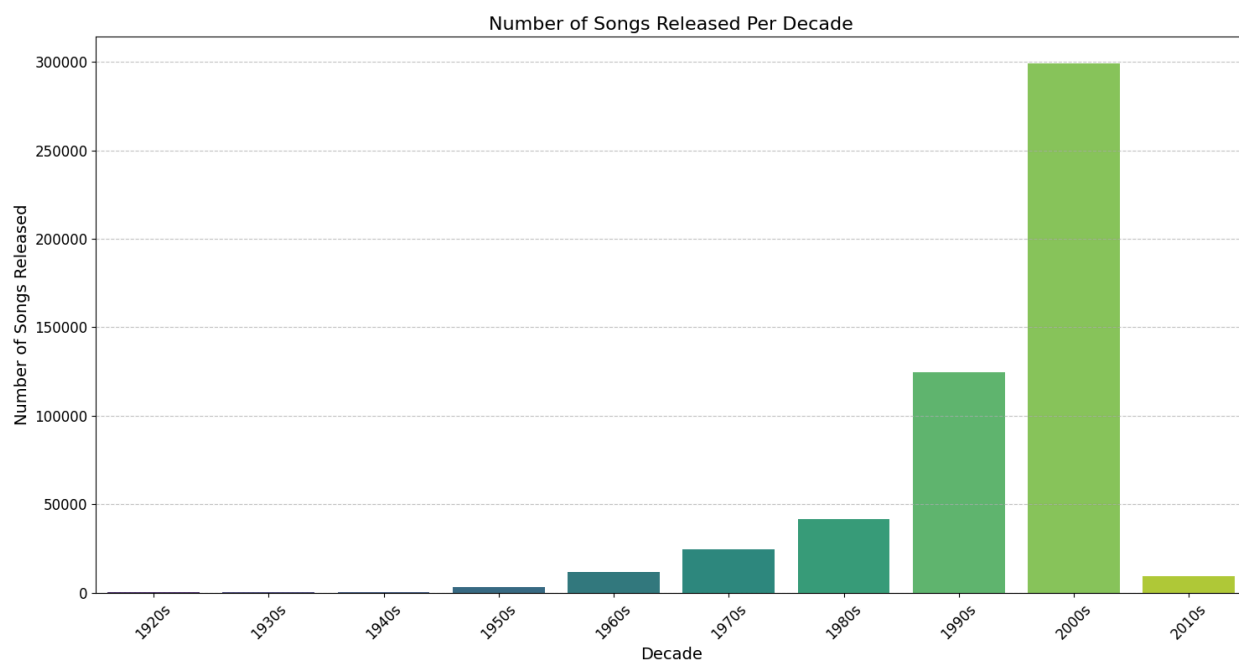

Business Recommendations and Next Steps

In order to increase website traffic and increase each users' average listening times, it is imperative to suggest songs that will keep them listening for as long as possible. This not only increases user satisfaction with the music platform, but also gives the company more advertising opportunities per user, driving up revenue and helps to maintain or increase subscription fees. The longer users listen, the more profitable the service can be. More traffic is directly related to higher profits.

The business should see this recommendation system as a constant work-in-progress. Spotify is constantly collecting new data on users and, therefore, they should continue tuning the model to be more accurate and more likely to suggest songs users will like.

The models are just guidelines. The best business result will come from constantly re-tuning and re-running the models. This recommendation system should be viewed as constantly evolving, not a final solution for the company.
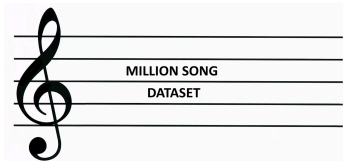
The results are not carved in stone. They are fluid and up for interpretation.

Spotify should hire more data scientists, pay them well to retain them, and train them to be as effective as possible. The business will fail without the recommendation systems. There is simply too much competition to ignore this fact.

Conclusion

User-user collaborative filtering performed decently well, but did not perform as well as the other models based on RMSE alone. Popularity recommendations are fairly ineffective compared to other models, but they perform best with new users who have not listened to any songs yet. This cold start problem can be handled with popularity based recommendations.

The item-item collaborative filtering and SVD models appeared to be the best recommendation models after tuning. The primary metric used to make this determination was the presence of low RMSE scores. This metric highlights the close relationship between data points within these models. This suggests these models in particular can recommend new songs to users with a high rate of similarity and relevance to users' individual tastes.

Many songs had a low or null play count from users. These missing and somewhat irrelevant songs leave gaps in data. Sparsity in the matrix means that SVD and co-clustering are generally more effective predictors than popularity and collaborative filtering. However, there is some nuance to this, as we have seen in the data.

The dataset lacks a dedicated rating system. This means that we have to predict the next song based on the users' play counts, rather than their actual subjective rating of a song. By eliminating songs with play counts greater than 5, we standardized the data and prevented skewed data that may arise from rating outliers.

The analysis proves a combination of two models to be the most effective approach to predicting play counts. Recommend popular content to new users

until their preferences become more clear as they create more data points for every song played. Once they hit a certain threshold we will switch to a Singular Value Decomposition model for that user. This is because of the relatively low RMSE score obtained through SVD. We can also save on computational costs by scaling the SVD model.

Appendix:

Item-item similarity was used to predict ten songs related to, "Learn to Fly." These songs were "Generator", "Stacked Actors", "Big Me", "X-Static", "For All The Cows", "Exhausted", "Wattershed", "Floaty", "Oh_ George", and "I'll Stick Around". The same process can be performed for any song in the cleaned database.

The Pareto chart shows the exponential growth of song releases per decade, and a sharp drop-off in the 2010s. There is not enough data to make up the entire decade of the 2010s, therefore releases are lower, but the gap is vast, so there must still be a change in the overall frequency of songs released.

Number of Songs Released Per Decade with Cumulative Percentage
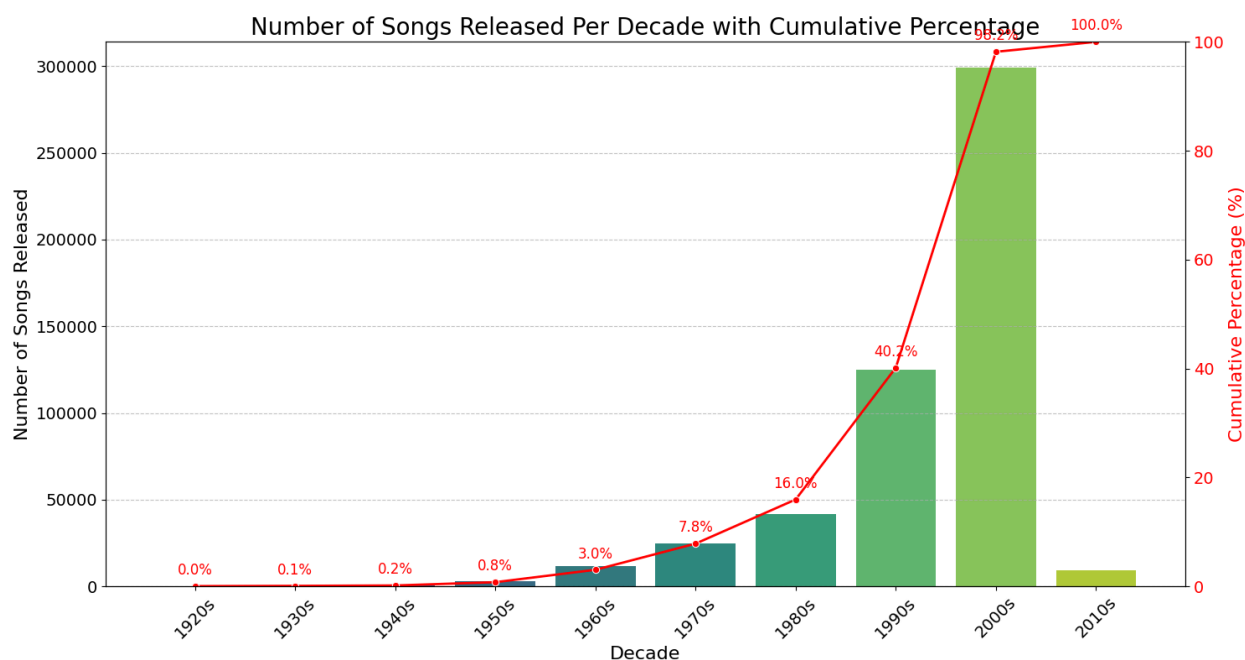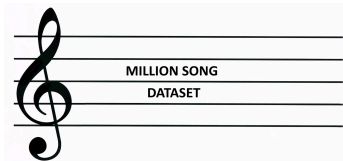
Citations:

- Hill, J., & Ashford, J. (2024, July 11). A brief history of music formats. UnifiedManufacturing. https://www.unifiedmanufacturing.com/blog/a-brief-history-of-music-formats/?srsltid=AfmBOopfytHL_EkHhPX1xOQvOaCOA1l0x7eMKCyCLyesc2FjhHaX3Ehn

- Welcome!. Million Song Dataset. (2011, February 8). http://millionsongdataset.com/

- Wikimedia Foundation. (2025, November 10). The Echo Nest. Wikipedia. https://en.wikipedia.org/wiki/The_Echo_Nest

- Wikimedia Foundation. (2026, January 2). *Singular value decomposition.* Wikipedia. https://en.wikipedia.org/wiki/Singular_value_decomposition