Music Recommendation with LFM-1b Dataset

Gasper Qian
University of Washington
xqia756@uw.edu

Dongyang Wang
University of Washington
dwang30@uw.edu

Wendan(Emily) Yan
University of Washington
emwyan@uw.edu

Abstract

In this project, we describe our design for several recommendation systems that help recommend music to users based on their personal background, such as gender, age, country, similar users, etc. The purpose is to utilize the recommendation systems such that the users can obtain music recommendations that best fit their tastes. Our models include collaborative filtering, demographic filtering, factorization machine, as well as hybrid filtering. We apply a 90-10 train-test split approach on the dataset and evaluate the models performances based on precision, recall, and the F-1 score. We find that across different number of neighbors, the factorization machine method has the best performance.

1 Introduction

Music is an integral part of human culture and has the power to evoke emotions, shape our moods and create memories. With the proliferation of digital music streaming services, such as Spotify, Apple Music, and YouTube Music, the number of available music tracks has increased exponentially. This has resulted in an overwhelming amount of choice for users, making it difficult for them to find new and relevant music. In recent years, music recommendation systems have emerged as an effective solution to this problem.

A music recommendation system is a software system that utilizes algorithms and data analysis to suggest music tracks to users based on their preferences, listening history, and demographic information. The goal of these systems is to provide personalized recommendations that meet the individual needs and tastes of each user. As a result, music recommendation systems have become an important area of research in the field of information retrieval and have been widely adopted by music streaming services.

According to a recent survey conducted by the Recording Industry Association of America, paid music streaming services have become the largest source of revenue for the music industry, accounting for 80% of all revenue in the industry. This has emphasized the importance of effective music recommendation systems and has motivated researchers to explore various approaches and techniques for building these systems.

In this paper, we first provide an overview of the current state-of-the-art in music recommendation systems and introduce the dataset. We start by reviewing literature on content-based filtering, collaborative filtering, and hybrid approaches, as well as documentations on the dataset itself. We then apply the various techniques used in these systems and proceed with our modeling in Python. Finally, we provide an overview of the evaluation metrics used to measure the performance of music recommendation systems we built and suggest future research directions.

Overall, this paper aims to provide a comprehensive understanding of the current state of music recommendation systems and the various approaches and techniques used in these systems. We believe that this paper will be useful for researchers and practitioners in the field of information retrieval and music recommendation systems.

2 Relevant Literature

There are two papers on the original dataset. Namely, they are *Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset* (2017) and *The LFM-1b Dataset for Music Retrieval and Recommendation* (2016).

In the two papers, Schedl presents the LFM-1b dataset. Also, the author conducted some Exploratory Data Analysis, generating a few key statistics such as the distribution of listening events over weekdays, hours of day when the listening events occurred. There has been shown to be a long tail of artists, meaning that most popular artists were played way more times per user compared to a large number of the least popular artists. In the 2017 paper, the author further incorporated the country level music taste into the analysis, showing that the countries the users came from were determinant of their tastes. The papers at last suggested a few recommendation systems that could be used in music retrieval and recommendation. Here is a list: collaborative filtering, demographic filtering, content-based filtering, hybrid recommender, popularity-based recommendation, and random baselines. The best performance comes from the collaborative filtering and the hybrid recommender, in terms of precision and recall, based on results from 10 fold cross validation.

Large-scale Analysis of Group-specific Music Genre Taste From Collaborative Tags (2017) describes the LFM-1b User Genre Forifile dataset which is the extension of the LFM-1b. This annotates listening events on the user level. It is related to the recommender system topic that is introduced in class along with the originated papers and recommender system algorithms that are used on LFM-1b dataset.

The dataset provides aggregated information on LFM-1b at a higher level, i.e. genre, and conducts comprehensive analysis, especially on correlations between musical preferences on the genre level. Statistical analyses were conducted to obtain insight into the distribution of listening events over genres per user group, the consistency of genre preferences within groups, and correlations between genres. Evaluation matrices are designed with two indices: Measuring agreement by Krippendorff's α, and Pearson's correlation coefficient to measure the agreements between users against the same music genre preferences, and the correlations within and between each attribute, e.g. the correlation between music genres.

Listener Modeling and Context-Aware Music Recommendation Based on Country Archetypes (2021) introduces a recommender system algorithm which implements a variational autoencoder algorithm that outputs context-aware music recommendation based on country information; in contrast to most previous works, we consider user country in our approach without using any external information about the country (clusters), such as cultural, economic, or societal information.

Based on the known track history of a target user, the models generate a variational distribution. As a result, the algorithm identifies country clusters (similar preferences) and archetypes of music listening preferences, and Top-k track recommendations are then retrieved by ranking the mean values of this distribution.

3 Data Collection and Exploration

The dataset we use will be the LFM-1b dataset obtained from the referred page of the course project website. Namely, the dataset was contained in the paper *The LFM-1b Dataset for Music Retrieval and Recommendation* (2016).

3.1 Data Collection

As decribed by Schedl, the dataset contains more than 1 billion music listening events created by more than 120,000 users.Each listening event is characterized by artist, album, track name, and also a timestamp. The dataset is unique in terms of its substantial size and a wide range of additional user descriptors it contains..

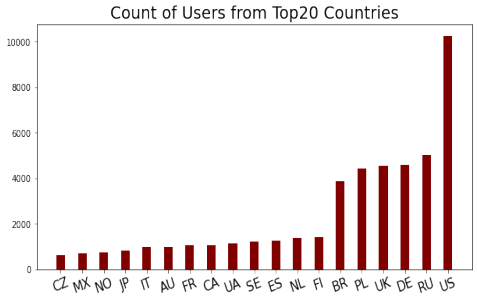
The dataset was obtained from the 250 top tags, which were later used to gather the top artists from Last.fm. The top fans for those artists were also obtained, from which a subset of users was taken. Among those users, their listening events were capped at 20,000 to avoid outliers. The dataset

contains “one billion music listening events created by more than 120,000 users of Last.fm. Each listening event is characterized by artist, album, and track name, and further includes a timestamp.

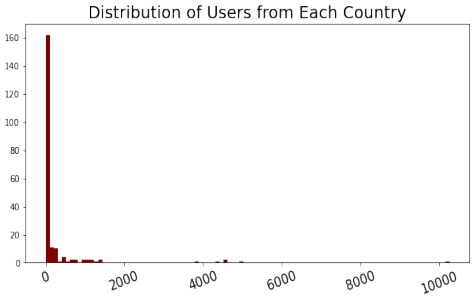
In the models, we primarily used the user artist playcount matrix, which contains a sparse matrix of listening events, where each pair of user and artist displays a playcount.

3.2 Data Exploration

We have performed some basic data exploration to visualize the key variables. We first illustrate the country of origin of the users. We can see that among the top 20 countries, the US, Russia, Denmark, the UK Poland and Brazil have the most number of users. Each of them have more than twice of the number of users from each of the remaining countries. Figure 1.2 shows the distribution of count of users from each country and we can see that this is a very skewed distribution with most of countries have fewer than 2000 users.



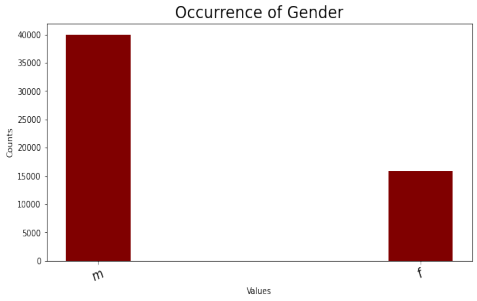
(a) figure 1.1



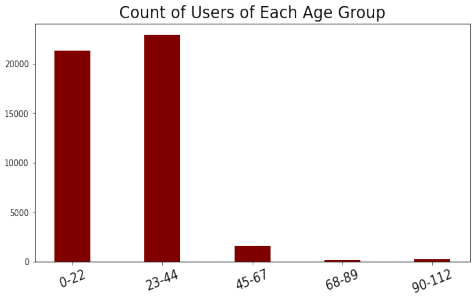
(b) figure 1.2

Figure 1: Country of Users

The figure below shows the gender and age of the users. We can see that male listeners are over twice of many as the female listeners so there is strong gender imbalance in the data. Figure 2.2 shows that most of the users are from the young adult group (0-22) and the mid-age group(23-44).



(a) figure 2.1



(b) figure 2.2

Figure 2: Gender and Age of Users

Figure 3 shows the number of play counts by different quantiles for each user-artist pair. We only shows the 0-90th quantile since above the 90th quantile the play counts have a significant jump and it’s hard to plot the scale on the same plot. From this plot we can see that for most users and most artists, the play counts is very small. This means that for the majority of artists, they are not listened to very much by the users.

4 Modeling

For our analysis, we have modeled the baseline collaborative filtering, collaborative filtering with only playcounts, collaborative filtering with playcounts and demographic information, demographic filtering, and hybrid filtering.

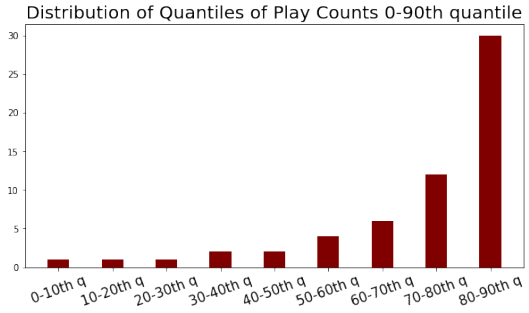


Figure 3: Distribution of Play Counts by Each User

4.1 Baseline Collaborative Filtering

For the users in the sample, we have emulated Schedl's approach with collaborative filtering. This approach recommends the top k nearest neighbors of the target user using the similarity measure of Pearson correlation coefficient. The formula is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

In the above formula, x_i and y_i represent the playcounts provided by the two users, and $\bar{x}$ and $\bar{y}$ represent the average playcounts of each individual user. The baseline approach collects all possible artists and directly returns the recommendations in numerically ascending order. This will simply serve as a baseline, since it contains no empirical meaning.

4.2 Collaborative Filtering with Playcount

Building up on our collaborative filtering model, we further improve on its performance by sorting the playcounts for the recommended artists. The recommendation will therefore be made based on the neighbors' frequency of listening. So, we have sorted the output in descending order of the total frequency of all artists listened by the user's neighbors and provided the recommendations accordingly.

4.3 Factorization Machine

Factorization Machine(FM) is a generic supervised learning model that maps real-valued features into a low-dimensional latent factor space. The FM model represents user-item interactions as tuples of real-valued feature vectors and numeric target variables. For each row of features, we identify the user in the user columns as binary indicator variables, and we identify the item in the item columns as binary indicator variables(see in Figure 5). We can also include auxiliary features for the users (such as age, country, sex, etc) and features for the items(genre, company, color, etc), as well as the context of the interaction (time, platform, link, etc).

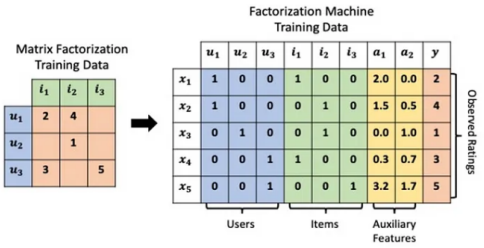


Figure 4: FM Dataset

The key design of the FM that sets it apart from a regular linear regression model is that it uses factorized interaction parameters: the weights for the interaction terms in the feature are the inner product of the two features' latent factor space embeddings. This way we reduce the number of parameters to estimate and at the same time break of restrictions on the independence of features. The optimization equation is below:

$$f(x) = w_0 + \sum_{p=1}^{P}(w_p x_p) + \sum_{p=1}^{P-1}\sum_{q=p+1}^{P}(< v_p, v_q > x_o x_q)$$

Another key design of FM is that it uses Learning-to-Rank (LTR) instead of the training error. The loss functions are based on the relative ordering of items instead of their raw scores. The LTR approach that we chose for our experiment is Bayesian Personalized Ranking (BPR). BPR maximizes the posterior probability of the model parameters given the observed user-item preferences and the prior distribution of the model parameters. In other words, we are finding the parameter θ by maximizing the below equation:

$$\theta = Max_\theta ln \prod_{(u,i,j)inS} \sigma[f(u,i|\theta) - f(u,j|\theta]p(\theta))$$

Where u stands for a user, i stands for an observed item, and j stands for an un-observed item. In our experiment, we adopted two methods to give a target user recommendations of artist. One approach is similar to collaborative filtering and demographic filtering: we find the K nearest neighbors of the target user and make recommendations the same way as the collaborative filtering. The second approach is we directly use FM to make recommendations for the target user by choosing items that give the highest prediction score by f. The evaluation score for both methods could be found in the Evaluation section.

4.4 Demographic Filtering

Demographic filtering takes into account the preliminary findings we obtained in the Data section. we incorporate demographic filtering as an additional approach to recommending artists to users.

Specifically, we consider users' gender, age, and country of residence to define a user similarity matrix. Using this matrix, we identify the K most similar users to the target user based on their demographic characteristics. To determine demographic similarity, we use binary criteria for gender (1 if same gender, 0 otherwise) and graded criteria for age and country (e.g., 0.8 if the age difference is between 1 and 2 years, 0.5 if the age difference is between 2 and 9 years, 0.2 if the age difference is between 9 and 15 years; 1 if the users reside in the same country, 0 otherwise).

We then combine these three similarity functions linearly, giving equal weights to all components. Finally, we perform aggregations through the collaborative filtering approach to recommend the top N artists to the target user.

4.5 Hybrid Filtering

We also propose a hybrid recommendation system that integrates both demographic filtering and collaborative filtering approaches. We adopt a late fusion strategy, where we fuse the results of the two recommendations using median normalization.

Specifically, we first normalize the ranking scores given by each system based on the median value of the scores. Next, for artists that are suggested by both recommendations, we compute the new score as the arithmetic mean of both original scores. For all other artists, we take the original normalized scores. This process helps to balance the contributions of each recommendation and improve the quality of recommendations. Based on the ranking obtained by sorting with respect to the new scores, we then recommend the top N artists to the user.

This hybrid recommendation system has the potential to provide more accurate and diverse recommendations to users, and we anticipate that it will perform better than either individual recommendation system.

5 Evaluation, Conclusion and Future Work

5.1 Evaluation

With the dataset, we perform a train-test split approach. Namely, we will mask 10% of the users' listening history to reserve as the test set to evaluate the performance of our models. After that, we perform evaluations based on the precision, recall, and F-1 score, which are all based on the confusion matrix shown below.

		Predicted	
		Recommended	**Not Recommended**
Actual	**Listened**	True Positive (TP)	False Negative (FN)
	Not Listened	False Positive (FP)	True Negative (TN)

To be more specific, the metrics we used are

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Also, for the convenience of computation, we have randomly selected 2,000 users to perform analysis upon.

Out of the masked listening history, we then perform an evaluation based on precision, recall, and F-1 score. We slice the number of recommendation equivalent to the number of masked artists for each user, and calculate their true positives, false negatives, and false positives. Subsequently, we can generate the desired statistics.

We have done the above procedure for multiple number of neighbors (K) values, and plot out the distribution of the metrics while varying K (figure 5). The Figure 6 shows the performance when we directly FM to recommend different number of artists. We can see that as we increase the number of recommended artists, the precision decreases and then plateaued, the recall increased and then plateaued. The F-1 score increased and then plateaued.

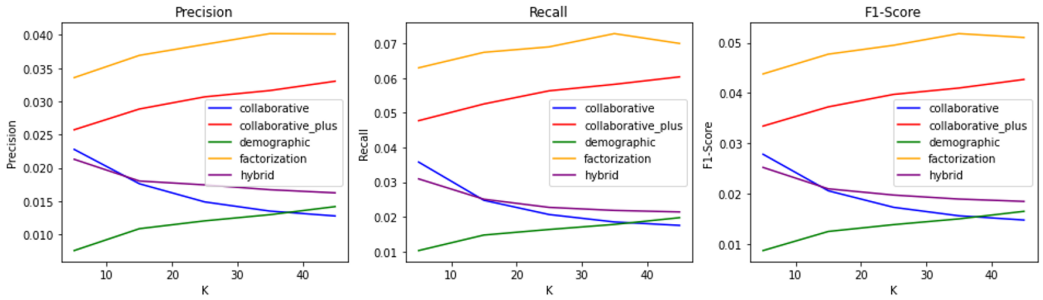


Figure 5: Precision, Recall, F1 Score for Different K Across Models

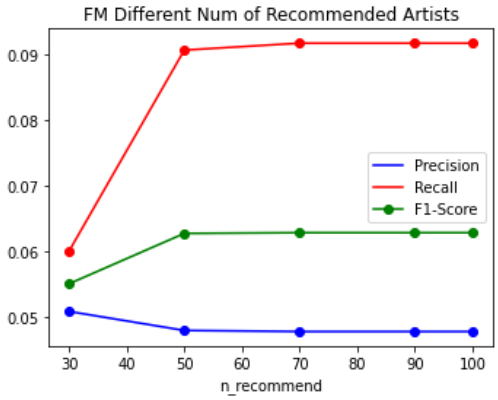


Figure 6: Precision, Recall, F1 Score for Different K Across Models

5.2 Conclusion

In conclusion, this study explored the use of various recommendation algorithms on the LFM-1b dataset. Our results indicate that factorization machine performs the best in terms of precision, recall, and F-1 score.

However, our study also highlights the importance of carefully selecting the evaluation metrics and dataset splits to ensure fair and meaningful comparisons between different algorithms. We observed significant variations in performance depending on the choice of evaluation metric and dataset split, especially the choice of number of neighbors K.

Despite the promising results, our study is not without limitations. The LFM-1b dataset is relatively small and focused on a specific music domain, which may not fully capture the complexity and diversity of real-world recommendation scenarios. Future research could explore the use of larger and more diverse datasets, as well as more advanced techniques such as deep learning models.

In practical terms, our research has implications for the development of recommendation systems in various domains, including music, movies, and e-commerce. By leveraging the power of various recommendation systems, businesses can provide more personalized and accurate recommendations to their users, leading to increased user satisfaction and engagement.

Overall, our study contributes to the growing body of research on recommendation systems and highlights the importance of rigorous evaluation and the exploration of new techniques to improve the performance of these systems.

5.3 Future Work

There are way more to do with the recommendation systems. For one thing, we can keep improving on more nuanced adjustments of the values of neighbors K, as well as the number of recommendations

to give. In these ways, the performance can probably improve and the recommendations based on these models are more likely to be in effect.

It's also possible that adjusting the weight between collaborative filtering and demographic filtering in the hybrid model will make the hybrid model perform better. Due to time constraints, this possibility has not been fully explored.

In terms of evaluation, another approach would be to mask 10% of the users off as the validation or testing set. In this way, we enable the models to predict unseen users' tastes. This approach would be valuable in real-world scenarios where new users start generating data and we would like to make recommendations.

To tailor our model to the best application in the real-world music streaming industry, we also considered the following model improvement areas:

1. Make recommendations based on popularity of artists. The amount of data we have for popular and mainstream artists differs from that of a more niche market. We hope to identify the difference between these two different markets and make recommendations that are unique to each group. This means we may need to handle the sparse parts of the user-artist matrix differently.

2. Considering the application of music recommendations, we think to recommend tracks instead of artists will be more useful in real-life situations. We will explore track recommendations instead of artist recommendation.

6 Individual Contributions

6.1 Gasper

Implemented methods on data retrieval, data pre-processing including normalization and imputations on missing data, randomized data sampling. Implemented demographic filtering and hybrid filtering along with result aggregation and demonstration related program. Write part of lit review, data collection, modelling, evaluation according to the associated implementations above.

6.2 Dongyang

Create the skeleton of the code and paper. Organize meetings and maintain meeting logs. Code/Edit collaborative filtering and introduce playcounts. Test code and write part of evaluation code. Write introduction, part of lit review, data collection, evaluation, conclusion, and reference parts.

6.3 Wendan

Coding the EDA part. Method research, coding for the factorization machine part, model training and tunning, and main part of evaluation, coding for the visualizations and plots. Write the EDA, Factorization Machine, Evaluation and part of the next step in the paper.

References

[1] Schedl, M. (2021, February 2). *Front. Artif. Intell.Sec. Machine Learning and Artificial Intelligence* Volume 3 - 2020 | https://doi.org/10.3389/frai.2020.508725

[2] Schedl, M. Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. *Int J Multimed Info Retr 6, 71–84 (2017).* https://doi.org/10.1007/s13735-017-0118-y

[3] Schedl, M. (2017). Large-scale Analysis of Group-specific Music Genre Taste From Collaborative Tags. *Journal of Music and Technology*, 5(1), 1-14. https://doi.org/10.1145/3078072.3078076

[4] Schedl, M. (2016, June). The lfm-1b dataset for music retrieval and recommendation. *In Proceedings of the 2016 ACM on international conference on multimedia retrieval* (pp. 103-110).

[5] Stackexchange. (n.d.). 2x2 confusion matrix. [online forum post]. Retrieved on 11th February 2023 from ttps://tex.stackexchange.com/questions/505915/2x2-confusion-matrix

[6] The Verge. (2019, September 6). Streaming revenue drives growth in the music industry. Retrieved from https://www.theverge.com/2019/9/6/20852568/streaming-revenue-growth-spotify-apple-music-industry-ariana-grande-drake-taylor-swift