Artist-Based Music Recommender

Springboard Capstone Project 1

Nicholas Taub

April, 2018

The Problem

The digital music industry faces increasing challenges with the rapid emergence of on-demand audio streaming. While online music streaming has proven itself successful as the preferred method for personal audio consumption, the expansion of the music streaming space poses fundamental issues for streaming's reputation as informative and resourceful. Users continue to display aligned listening habits, however the proliferation of available music undermines the opportunities for like users to "discover" new music that they would've never enjoyed without attention to the similar preferences of fellow listeners. Users should expect to experience creative limitations, that is, constraints on their personal music library, if streaming services do not recommend music which fosters musical exploration by individual users.

The Situation

Among popular audio streaming providers, rising usage rates fuel the creation of diverse user and item data. As users oscillate between their artists of choice, companies obtain vast informational insights into the patterns derived from music listenership. But how do these insights help broaden user listening habits? What application features might foster artist discovery? While audio streaming is useful for celebrating listening *habits*, user listening *preferences* are often underestimated, thus limiting opportunity to make recommendations to users as to which other artists they might enjoy based on their similar preferences.

The Solution

Item-based collaborative filtering would effectively promote scalability of the recommendation engine given the cold start problem that would quickly arise due to rapid registration rates. Equivalent to the traditional music retail associate, but with the advantage of knowing other like customers, collaborative filtering would initially identify shared artist preferences across a broader user base, and then predict unheard artists based on like artists previously listened to by other users. This logic rests on the key assumption that past listening similarities between artists should determine those items' similarities over time.

The Client

On-demand music streaming service with a focus on accurate and scalable recommendation systems.

Data Collection

The public datasets collected for this report were accessed online from Òscar Celma at the Universitat Pompeu Fabra, one table including user profile information, the second containing playcount per user by artist. The data includes approximately 358,868 users, 292,364 artists comprising a total of 17,535,655 listening observations.

Raw features of the profile data are user ID, signup date, gender, age and country. The activity data comprises of user ID, artist, artist ID and playcount.

Data Wrangling

Within the profile data, missing age values were populated with the existing average age. The profile data was further filtered to only include user within the age range of 15 to 45 years old. Limiting the age range promotes thorough listenership for feature inputting while reducing low-end and high-end outliers.

The profile and activity tables were then merged on user IDs. All rows with missing values were dropped. The data was then filtered to includes users with a minimum selection of 20 unique artists. Artists with a total playcount below 50,000 were subsequently excluded too.

After cleaning, the profile dataframe included entries for 26,306 users and 7,668 artists:

Unique Users	Unique Artists	Unique Entries
26,306	7,668	1,192,596

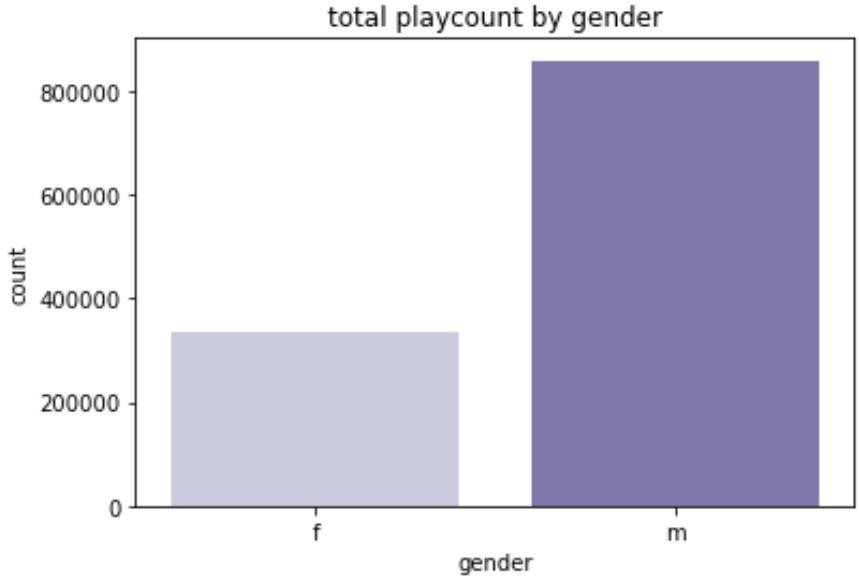
The userArtist dataset serves as the primary informational structure throughout the duration of this report. First, the dataset will assist with conducting EDA to visually assess the relationship between demographics and listening patterns. Inferential statistics will thereafter test the validity of spetributes of the user data. Finally the data will be converted into a sparse representation of artist-artist convergences for implementation of item-based collaborative filtering.

Exploratory Data Analysis

EDA for this report focused on the influence of age and gender on user playcounts. Initial findings indicate that gender serves as a significant driver for user listenership. Age plays a pivotal role for determining playcount too, but as a continuous variable, as will be shown, its influence can be limited over time.

Gender-Specific

To begin, below is a bar chart constructed illustrating the distribution of total playcount per gender:



The above visualization displays a discrepancy between male and female listening trends, with male total playcount more than double the total playcount of females. Interestingly, though males register at a greater rate than potential female users, the genders share relatively similar average playcounts as displayed in the chart below.

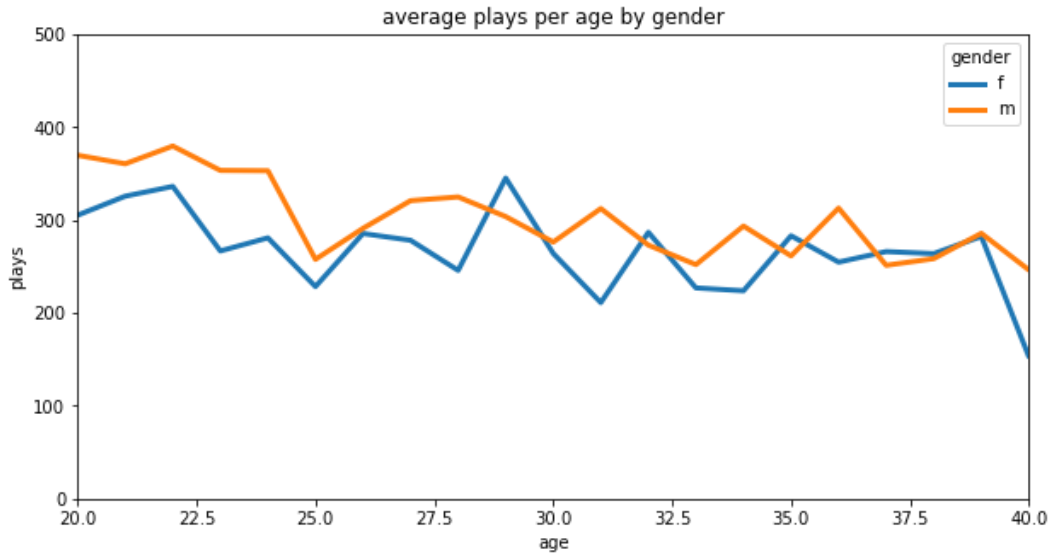
Gender	**Average Playcount**
Female	277
Male	305
Mean Difference	28

One natural driver for higher male listening activity might be associations with male artists in lieu of equally represented female artists. One source observes that the outnumbering of male artists in the music industry relates to organizational shortfalls to sufficiently foster female artistry.[1] This suggestion holds fairly true for other entertainment fields, like film, thereby reinforcing the notion that gender associations continue to thrive in favor of a predominantly male music scene.

[1] Martha Tesema, *Mashable*, Dec 2017, accessible at: https://mashable.com/2017/12/07/spotify-top-five-2017-all-men/#42L1DcHq_iqU

Age-Specific

The line graph below aims to place gender listenership in the context of the active age range of 20 to 40 years old. While gender distribution sways male, the difference of average playcount between genders is fairly narrow. Ever so slightly the playcount across genders begins to decline.



Statistical Inferences

This section aims to determine if the patterns within the sample data occurred by chance. The prevalence of male listening activity will be specifically measured for its application to a real population environment. Below are the hypotheses for conducting this test.

Hypothesis H0	Male and female playcount means are equal
Hypothesis H_a	Male and female playcount means are not equal

Null Hypothesis H_0 - male and female playcount averages have equal mean values

Alternative Hypothesis H_a - male and female playcount averages have different mean values

In light of the difference between average female and male playcounts, both means were measured with a 2-sample t test to confirm their representation of the populations from which they were taken. Though the sample data does not display a normal distribution, the sample sizes were large enough to rely on the *p value* to answer whether the null hypothesis was true or false.

Moreover, a significance level, α, was set at 0.05, or 5%, as the acceptable threshold for rejecting the null hypothesis. A <5% p value would warrant rejection of the null hypothesis and that that our sample data represents true differences in playcounts between men and women in a true population.

The SciPy t test function generated a p value of 6.96e-08, confirming a statistically significant difference in the female and male sample sets. The confidence level of this difference was 99.99999993%, concluding that the means observed in the gender-based sample data closely reflect the means in the greater population.

Prediction Methodology

With data prepared to only include active users and artists, a pivot table was generated to capture play convergences between users and artists. The pivot table was initially converted into binary values prior to reformatting as a CSR matrix. Initial findings of both user-based or item-based pairwise distances confirm relationships across both axes.

Collaborative filtering will derive from explicit, binary-based feedback whereby identifying items that are similar to the items a user has previously selected based on playcount. Binary inputs will allow for easier computation of cosine distance when measuring from -1 to 1.

Collaborative Filtering

Collaborative filtering for item-based recommendations leaned on scikit-learn's NearestNeighbors model using cosine distance as a key tuning parameter. When passed a random artist, the model was tasked with finding nearest distances of 6 neighbors within the sparse matrix created previously. Thanks to earlier efforts to reduce zero values, the model exhibited trusting results given the relevancy of the model's artist recommendations.

The success of item-item collaborative filtering demonstrates that genre is not necessary for all music recommendation systems. While genre would effectively optimize model precision, the simple task of cosine distance measurement serves efficient for pinpointing similar items per user.

Conclusion

This report assessed the demographic and activity patterns of a wide range of music data. The EDA illustrated a tendency towards male-driven registration regardless of nearly equal engagement of men and women regarding average playcount. Music recommender applications should thus investigate the potential factors driving male registration rates beyond that of females.

Regarding collaborative filtering conducted for this report, item-item is a reliable model when working with less artists than users. In this contect, artists are likely to present more feedback than you'd find when relying on user activity, often returned to as the "cold start" problem with user-based recommendation engines.

Further research would benefit from measuring the accuracy difference when inputting playcount as implicit feedback in the form of exact plays per artist by each user. Providing unique weights for each artist across users may encourage solution finding for the cold start problem mentioned previously. Additional research into the relationship of age and genre may also serve fruitful for identifying methods for recommender optimization.