BUILDING A MUSIC RECOMMENDATION SYSTEM

Project proposal for the requirements of the course

BANA 8083 MS Capstone

By

Piyush Verma (M12396911)

MS Business Analytics Candidate 2018

"On my honor, I have neither given nor received unauthorized aid in completing this academic work. This work has not been published/reported or submitted to any other university or institution for the award of any degree."

Department of Operations, Business Analytics & Information Systems

Carl H Lindner School of Business

University of Cincinnati

ABSTRACT

Streaming music have become one of the top sources of entertainment for millennials. Because of globalization people all around the world are now able to access different kinds of music. The global recorded music industry is worth $15.7 billion and is growing at 6% as per 2016. Digital music is responsible for driving 50% of those sales. There are 112 million paid subscribers for the streaming business and roughly a total of 250 million users, if we include those who don't pay. Thus, it becomes very important for streaming service providers like Youtube, Spotify and Pandora to continuously improve their service to the users.

Recommendation Systems are one such information retrieval technique to predict the ratings or popularity a user would give/have for an item. In this project I would be exploring bunch of methods to predict ratings of users for different artists using GroupLen's Last.FM dataset.

DATA

The data for the project is available free to use for non-commercial purposes here by GroupLens. GroupLens is a research lab at the University of Minnesota which publish research articles in conferences and journals primarily in the field of computer science, but also in other fields including psychology, sociology, and medicine. The owners constantly update the dataset in the page. At the time of writing this report this dataset contained social networking, tagging, and music artist listening information from a set of 2K users from Last.fm online music streaming site. The zip file contains namely the following files:

artists.dat	Contains information about music artists listened and tagged by the users
tags.dat	Contains the set of tags available in the dataset
users_artists.dat	Contains information on listening counts
user_friends.dat	Contains information on bi-directional user-friend relation
user_taggedartists.dat	Contains information on tag assignments of an artists provided by each particular user
user_taggedartists-timestamp.dat	Contains above information which the timestamp

Some detailed information about the data:

1. Information about 17632 artists tagged by 1892 users
2. An artist can be tagged differently by a different user
3. Weights in user_artists dataset is the listening count of an artist by a user

All the datasets would be used for making a recommendation system except user_taggedartists-timestamp.dat because it doesn't provide any new useful information. The number of time a user listened a particular artist will be used as a rating (or proxy for rating).

LITERATURE SURVEY

According to various blogs read, it is found that we can implement recommendation system in many ways: each with a different algorithm running them. Following are the potential methods:

1. **Clustering**: K-medoids form of K-means clustering can be used to group artists which are similar based on their genres. In this method, since all the predictor variables are categorical variables we can calculate "*gower distance*" between artists and perform clustering.
2. **Matrix factorization**: Here the idea is to approximate the whole rating matrix by the product of two lower dimension matrices. For this the idea is to predict the rating in such a way (using training dataset) such that the sum of squared error (=actual – predicted ratings) is minimum. Here we also regularize and introduce a penalty term β to avoid overfitting.
3. **Collaborative Filtering**: CF can be done in two forms:
    a. *User Based Filtering*: Where similar users are first found using their similar decisions in the past and then their items are recommended to each other. The rating prediction is done based on calculating the weighted average of user ratings and their similarity to the user in question. For calculating the similarity, the most popular choice is: centered-cosine-distance which easily differentiates a tough and an easy user.
    b. *Content Based Filtering:* Where if a user decides on an item, he/she is recommended the next item based on the item which is like the item he just decided upon. Here the predicted rating is calculated in a similar fashion as in user-based filtering.

In general, it has been reported that for many cases Content Based Filtering (also known as item-item) outperforms User Based Filtering (or user-user). And the reason is very interesting. Items are "simpler" than users in the sense that items can have only certain genres, but a user's taste can vary largely. For example, an artist can belong to metal rock but very unlikely that he/she would also belong to classical opera music. Where as users can like metal rock and also can listen to classical opera music occasionally.

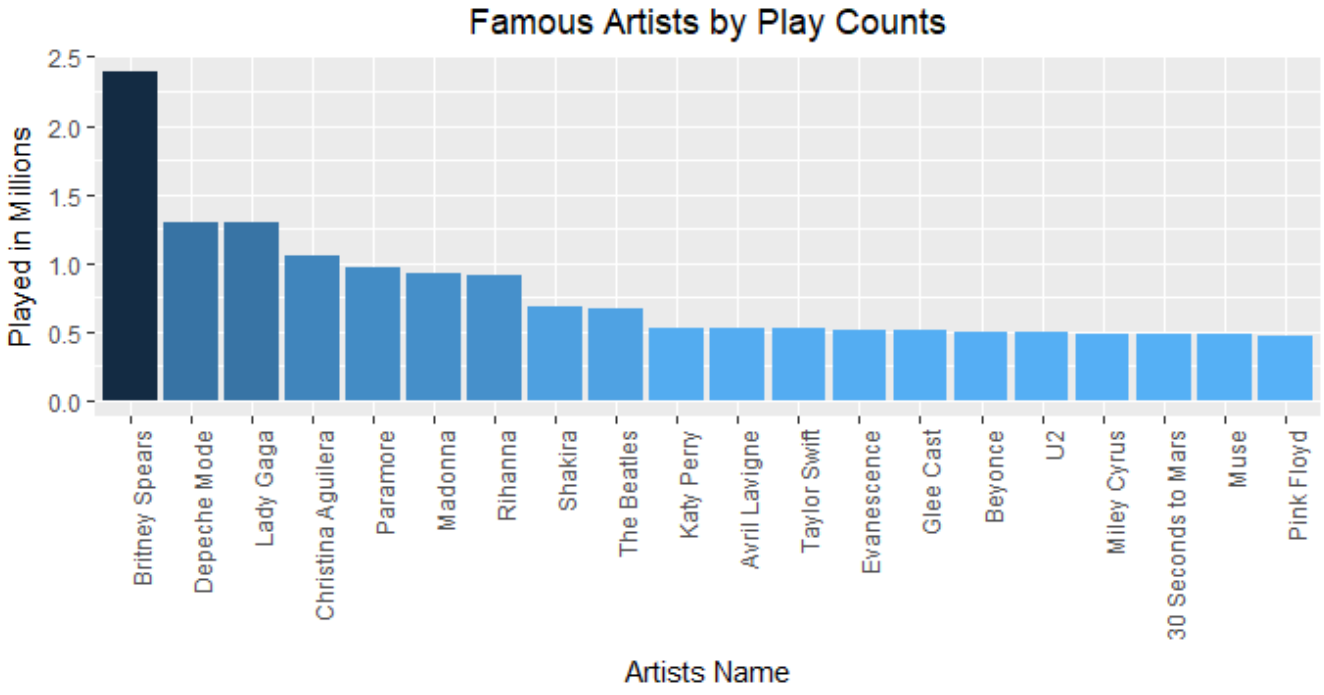
PROPOSAL

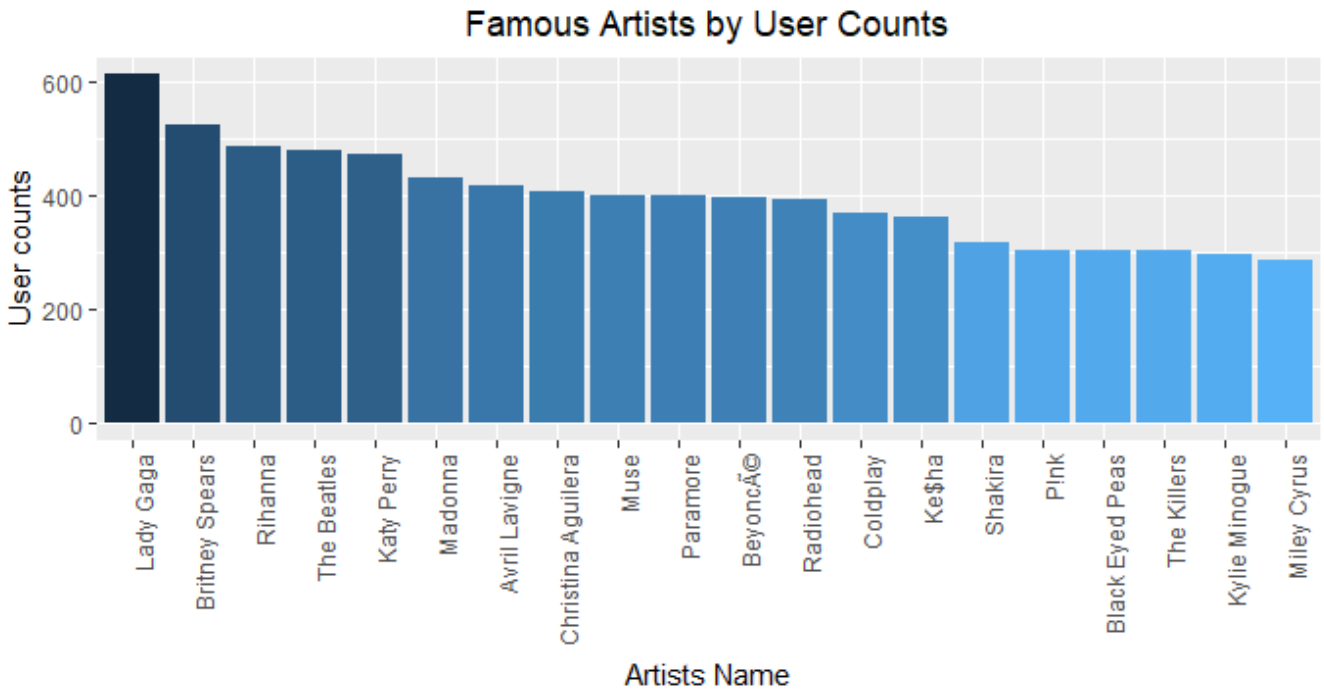
To implement a good recommendation system, steps can be taken in the following order:

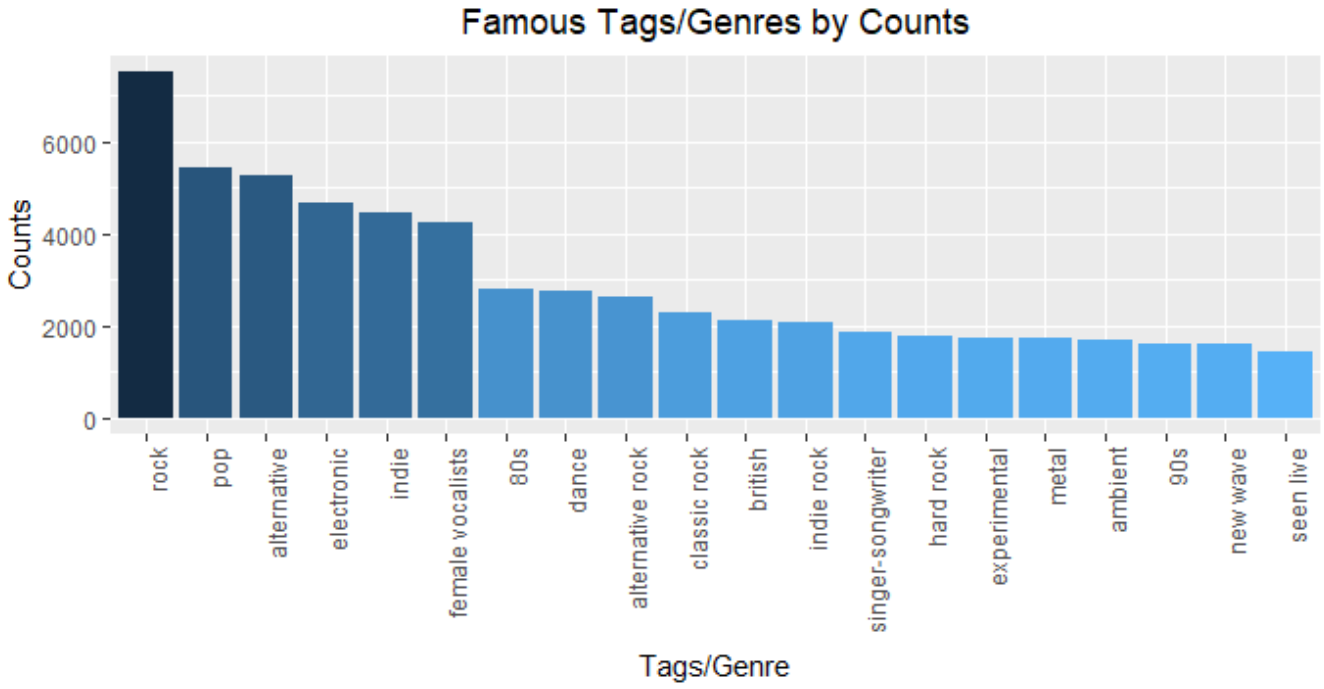
1. Clustering the artists based on “gower distance”, having 3 genres inputs from the user, finding the Top N artists for which the sum of “gower distance” is minimum
2. Using Matrix Factorization, predict the ratings for the artists which the user has not listened to and then recommending the Top N unheard artists
3. Using Content Based Filtering to calculate the weightage average ratings based on similarity and then recommending the Top N unheard artists
4. Comparison of the results from the above 3 methods

PRELIMINARY EXPLORTORY DATA ANALYSIS

To begin the analysis, EDA was performed to get an understanding of popular artists and genres in the dataset. Following were the results:







We can say that rock, pop and alternative are the popular genres. And among artists: Britney Spears and Lady Gaga are among the famous and most played artists.

BIBLIOGRAPHY

- Music industry (Abstract): https://www.forbes.com/sites/hughmcintyre/2017/04/25/the-global-music-industry-grew-by-6-in-2016-signalling-brighter-days-ahead/#33e56d6163e3
- Data Source: Grouplens.org http://files.grouplens.org/datasets/hetrec2011/hetrec2011-lastfm-2k.zip
- Role of Matrix Factorization Model in Collaborative Filtering Algorithm: A Survey https://arxiv.org/ftp/arxiv/papers/1503/1503.07475.pdf
- Matrix Factorization: http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/