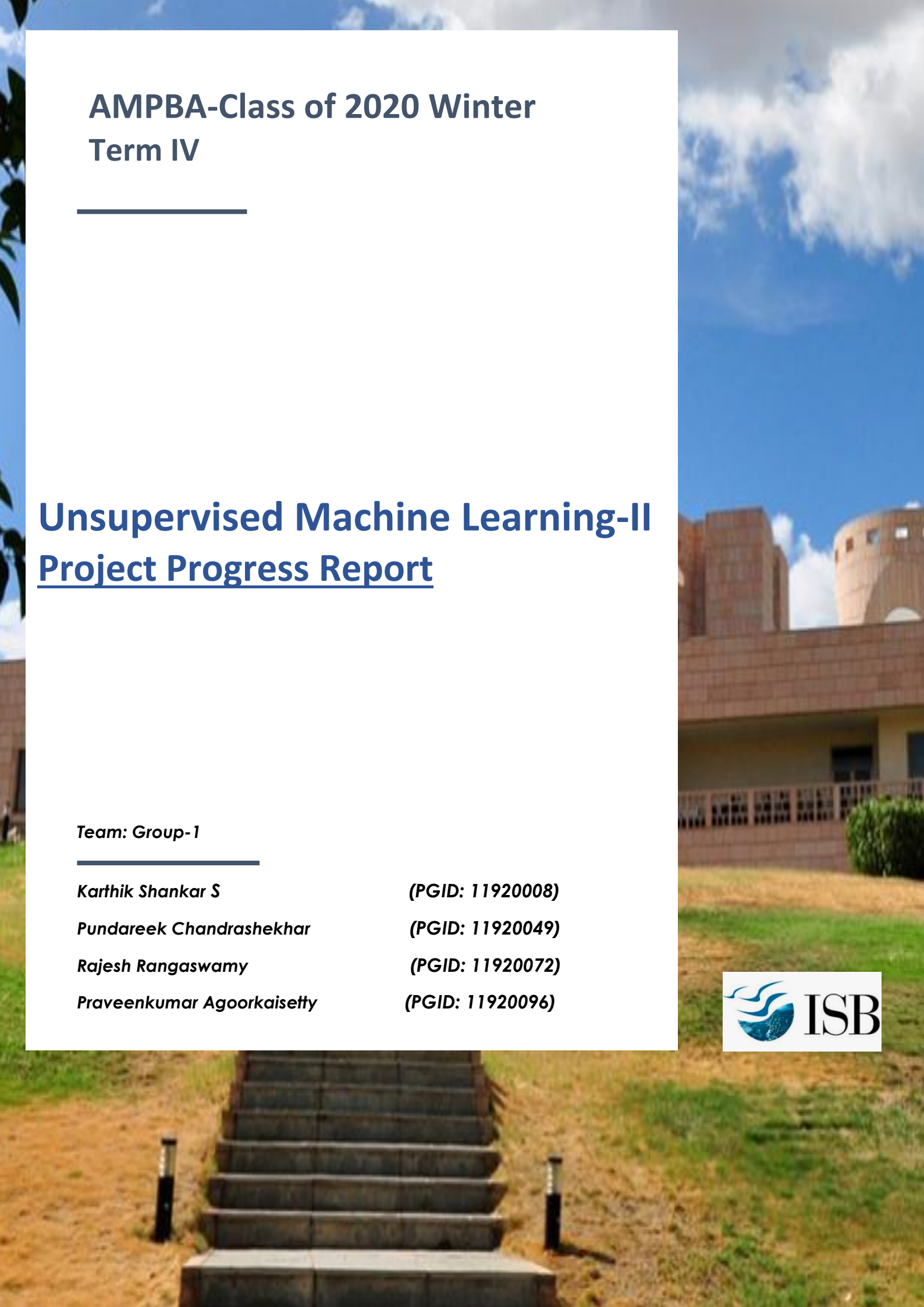


AMPBA-Class of 2020 Winter Term IV

Unsupervised Machine Learning-II
Project Progress Report

Team: Group-1

Karthik Shankar S ***(PGID: 11920008)***

Pundareek Chandrashekhar ***(PGID: 11920049)***

Rajesh Rangaswamy ***(PGID: 11920072)***

Praveenkumar Agoorkaisetty ***(PGID: 11920096)***

ISB

Project Goals and Objectives

Listening to music is obviously one's personal choice. But with the arrival of smart gadgets and availability of popular online entertainment channels, Preferential listening to music can now be coerced or an user can be recommended to listen a song based on their listening history in an entertainment ecosystem. By listening history means, a song listened by an user can be tagged to 'n' number of attributes for suggesting any other user with same or similar taste profile. The Project titled **"Song Recommendation System"** is an attempt to our Assignment Fulfilment for Unsupervised Learning-II as part of the Academic Curriculum.

Goal of the Project

To devise an Unsupervised Machine Learning model which would recommend a song to an user based on user taste profile(# of times a song has been played) based on collaborative filtering techniques. We will evaluate the model based on ROC curves which is reserved for final output. For achieving a desired result, we have reduced the size by only considering a subset and using only 1 feature first hand to achieve the desired working. Scaling of this by involving other features is just an extension of the project. The aim is to provide a realistic and relevant song recommendation based on user taste profile.

Objective of the Project

Objective of the attempt is to develop a Proof of Concept (PoC) for the Song Recommendation System by training and testing the model with the subset data first hand.

We aim to use machine learning to suggest or recommend a song based on both user and song characteristics with the metadata about a song (such as user listened and # of times listened) From the models, we hope to also get a deeper insight into the features that are most suggestive of a song popularity, and understand what makes certain songs more popular than others.

Our problem statement is thus

1. How to suggest or recommend a song based on user listening history?
2. What would be the song recommendation for a new user based on the model being developed?
3. What type of modelling technique would yield a better song recommendation or suggestion?

We propose to define success of this project by analysing the metric which would be obtained from the ROC Curves later.

Data Collection Steps

The main source of the dataset for the project in focus is referred from https://www.kaggle.com/c/msdchallenge and www.millionsongdataset.com.

The Million Song Data Set (MSDS) is a freely available collection of audio feature and one of the extensive database of popular contemporary tracks spanning decades of western music.

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 but only the derived features. Both the subset and the entire dataset are fairly large. The subset is 2.5GB while the full Million Songs dataset is 270GB. We initially tried building our models directly using the original dataset. However, that made it difficult to do data cleaning, imputation and statistical analysis. This has imposed the limitation on us to use the subset of the original dataset. Hardware requirements would be highly demanding for handling the bulk of dataset with even 3 to 4 key attributes for receiving a recommendation from the model. This actually would become a big data problem involving Hadoop and Map-Reduce. We have kept this issue to handle later and have decided not to make use of the entire dataset while we pursue on PoC objective at this point of time.

Hence, we switched to a different method based on the observation that a lot of the data in the dataset was not relevant to us at this instant (such as the actual audio tracks) so we filtered the data from original file and converted it into a much more compact CSV representation. The basic dataset with associated attributes are actually available in different files. Out of the 55 attributes associated to a particular song, only the required features were identified, grouped, extracted and then merged to get the workable subset over which the model will be run. We filtered the files by only extracting features relevant to our problem statement. For the 30,000 song subset, these included the following fields: "**song_id", "user id" and "# of times played"**. This filtering enabled us to convert the 270GB dataset into a significantly more manageable data file size.

Why this data suits to our Project Goals?

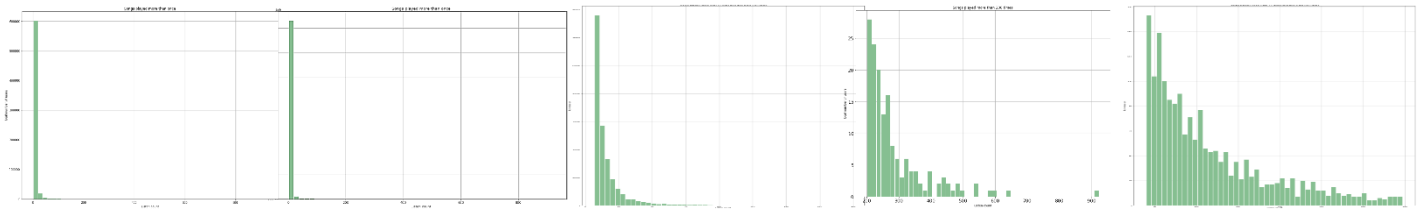
While a lot of work has been done on applying machine learning to different facets of the million song dataset, there has been little to no focus on song popularity. Research has focused on predicting the year, artist of a song, recommending songs based on genre etc. There is also a lot of existing literature on attempting to determine the characteristics of a popular song, but they have tended to focus more on evaluative or statistical approaches than a big data or machine learning approach. We present a system that ties together the two aspects we build machine learning models that are able to suggest a song based on listening history accurately and then use these models to get more insight into the features that are the strongest signals of song popularity.

Basic Analysis of Data (Descriptive Statistics, Visualization)

We have done a basic EDA on the whole of dataset to extract few details as below.

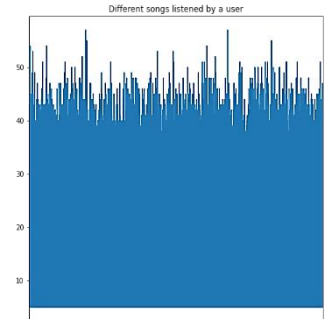
1.Total Number of times the song was played (Frequency of listening (X) versus Users (Y)).

To improve the visibility of Play Count versus number of times the song was played by the users, the scales have been progressively modified as below



Frequency 200 | Frequency >1 | 6>Frequency>200 | 75> Frequency>200 | Frequency >200

2.Different Songs listened by an User



Visualization of different songs listened by a user from the whole data set.

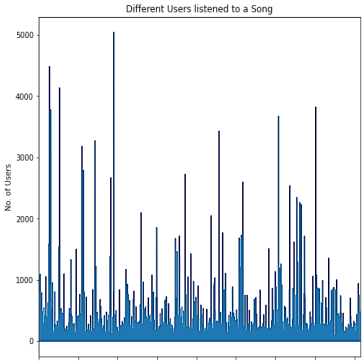
X-Axis: User ID

Y-Axis: # of Songs listened by the user

Purpose: To figure out how many songs are listened by an user.

We are interested in figuring out the minimum count of songs listened by an user. After this visualization, we are considering the user with minimum of 50 songs listening history in our model.

3.# of times a song was listened by unique users



Visualization of listening count of unique users for a particular song.

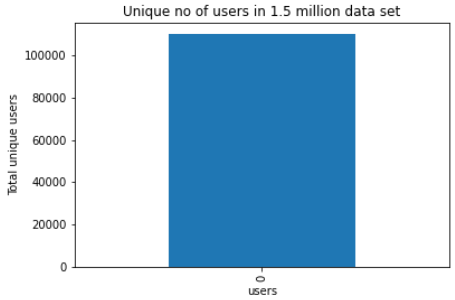
X-Axis: Song ID

Y-Axis: No of Users who have listened

Purpose: To figure out how many users have listened to the songs in the whole of repository.

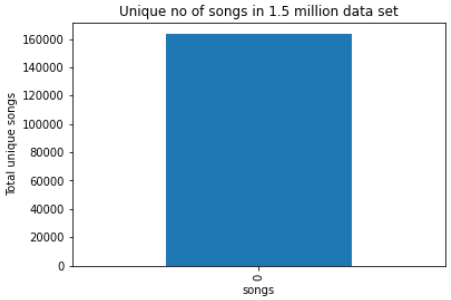
We are interested in using this information for recommending a song to an user when no listening history is avilable at first instance in our model. This information is helpful in formulating a criteria for a song to be recommended based on listening history by users, when we deal with Collaborative filtering.

4.Total No of Unique users in the Data Set



(ie.,110,000)

5. Total No of Unique Songs in the Dataset



(i.e.,163,206)

Preliminary Analysis

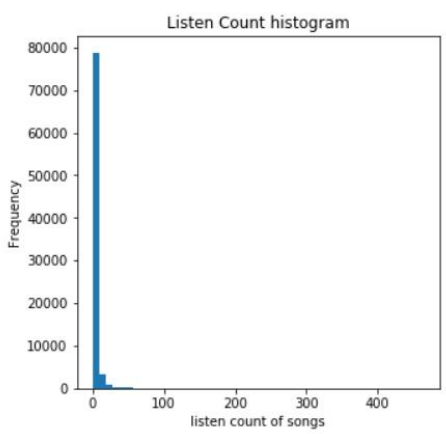
As mentioned earlier, we are limiting ourselves with only the subset of the whole data set with 30,000 (can be modified based on hardware resource availability) song records.

Limitation being, the model is very demanding on hardware requirements if we consider the bulk of dataset and we are afraid that, it may result in indefinite computation time in processing.

For the subset of the dataset (83,741 records), we have set the following filters

A song will be eligible for consideration only if a **minimum of 35 users** have listened to it.

User based recommendation based on user taste profile will be taken into account by the model only if the user have listened to **at least 50 songs** in the repository. They are defined as the active listeners in the eco-system whose user id will be taken into account for user based recommendation.



After the above filtering criteria

The listening count details are

Minimum Listen Count:1

Maximum Listen Count:463

A user who is new to this eco system would be recommended a song based on the popularity. (maximum user count for a song)

Description of the Way Forward

The final extracted dataset will be partitioned in the ratio of 80:20 for training and testing purpose respectively.

From here, we would build the model for song recommendation based on the Collaborative Filtering Technique mentioned as below.

1. User Based Recommendation
2. Item Based Recommendation

As the data is highly sparse, we would be using the following algorithms for calculating similarity for both the above mentioned collaborative filtering techniques.

1. Cosine Similarity
2. Correlation Proximity

After modelling is done using above algorithms, we would be using the test dataset for its performance measurement and would compare all the performance metrics.

Based on what we achieved, a final model would be selected and recommendation would be done.

Would see if an API could be built around the recommendation model for future usage.
