Song Recommender System

Dipika Sharma

Bellevue University

Applied Data Science 680

Amirfarrokh Iranitalab

Project White Paper: Song Recommender System

Introduction:

Using the machine learning to suggest the songs to the users that they might like, this system is called song recommender. As part of this project, I chose the dataset of track records. Using this analysis, we will recommend songs to users.

Business Problem:

Song recommender system will go through the user's history and create a list of songs as per their preferences. The intend of this system is to make users familiar with some new songs and album to improve user experience. This system will help business owners to grow their income as this system will encourage users to make more purchases by recommending music tracks based on their earlier purchase.

Data Explanation:

I will be using the dataset from Kaggle website.

The data can find at below location:

https://www.kaggle.com/datasets/subho117/music-recommendation-system-using-machine-learning/data

This dataset is good for our purpose as it consists of user's id and the purchases the users made in past. This dataset also consists of some scientific information about the songs like valence, acoustic Ness, and others.

The detail of the columns is below.

Valence	:	measure how positive or negative music sounds it measure on a scale of 0.0 to 1.0
Year	:	Year the music released
Acousticness	:	It measure on a scale of 0 to 1, it stat how much a song is acoustic.
Artists	:	It shows the artist of the song.
Danceability	:	Measure how easy to dance on a song.
Duration_ms	:	What is duration of song?
Energy	:	Used to measure the intensity of song.
Explicit	:	if the song contains any offensive content.
Id	:	The user id.
Intrumentalness	:	Measure on a scale of 0 to 1 and indicates if song is instrumental or not.
Key	:	scale of notes.
Liveness	:	If audience is presence in the song recording.
Loudness	:	Measure how loud or quiet the song is.
Mode	:	musical scale variation
Name	:	Name of the song.
Popularity	:	How popular the song is ?
Release_date	:	Release date of the song.
Speechiness	:	It measure the spoken words in song.

Tempo : It indicates the speed of the song.

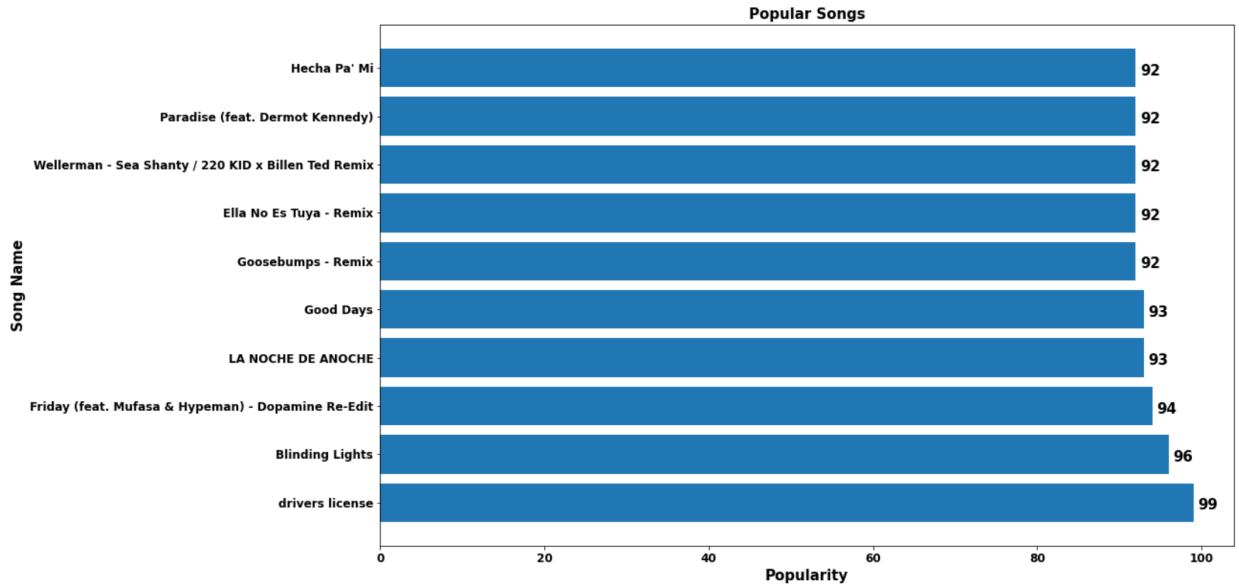
Project Methodology:

As mentioned, the objective of the project is to create a song recommender system to recommend songs to the user as per their past choices. A classic approach is to perform the analysis on tracks database which consisted of various song features and artist information and take help of music expertise to recommend the songs, but a more modern approach includes machine learning. I am using the cosine similarity. This approach is considered to be the best if we need to compare the two vectors to find out the similarity and to understand the relationship between the two-vector considering the relative presence or absence of musical attributes and prioritizes the similarity in sounds of the songs resulting in good recommendation This is appropriate approach to deal with tracks data and most widely known for efficiency, reliability, and capability when we need to build recommender system.

As part of this project the cosine similarity system is build and used to recommend the songs to the users.

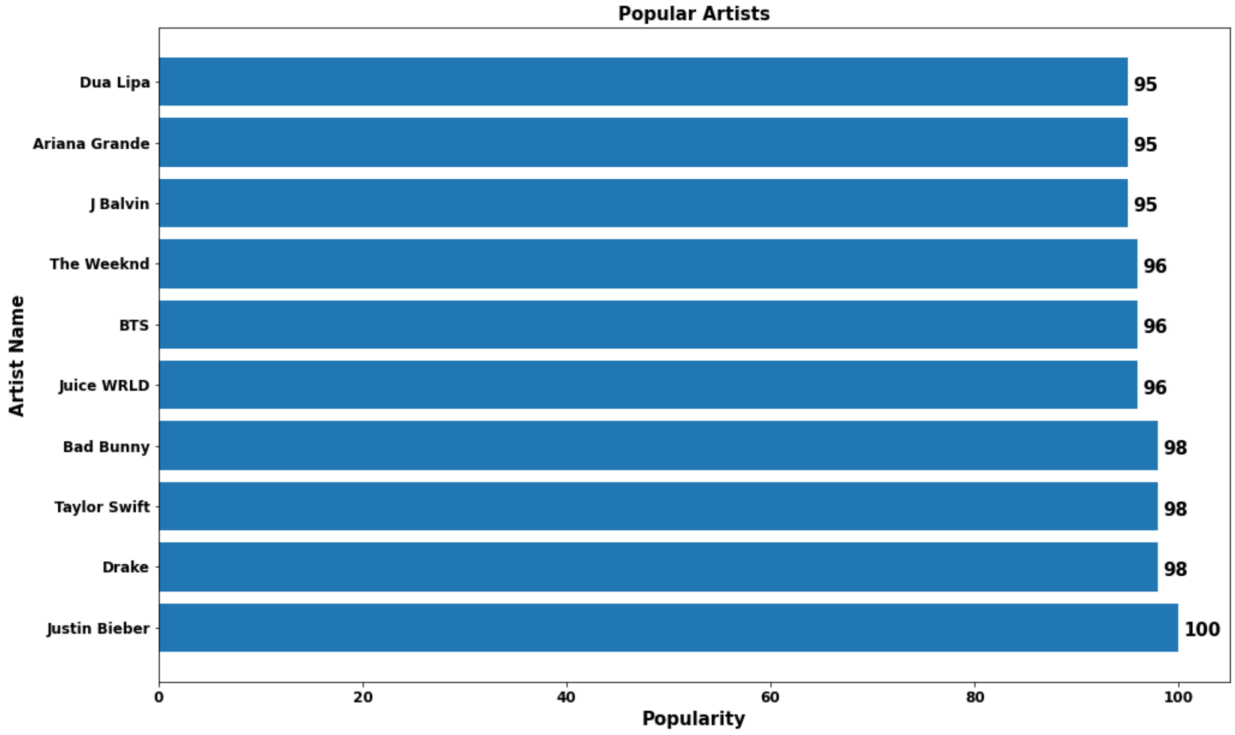
Analysis:

I started the project work with exploratory data analysis as this will help me understand the data. I used different visualization to draw some initial conclusion. As first visualization I analyze the top 10 most popular songs based on its popularity.



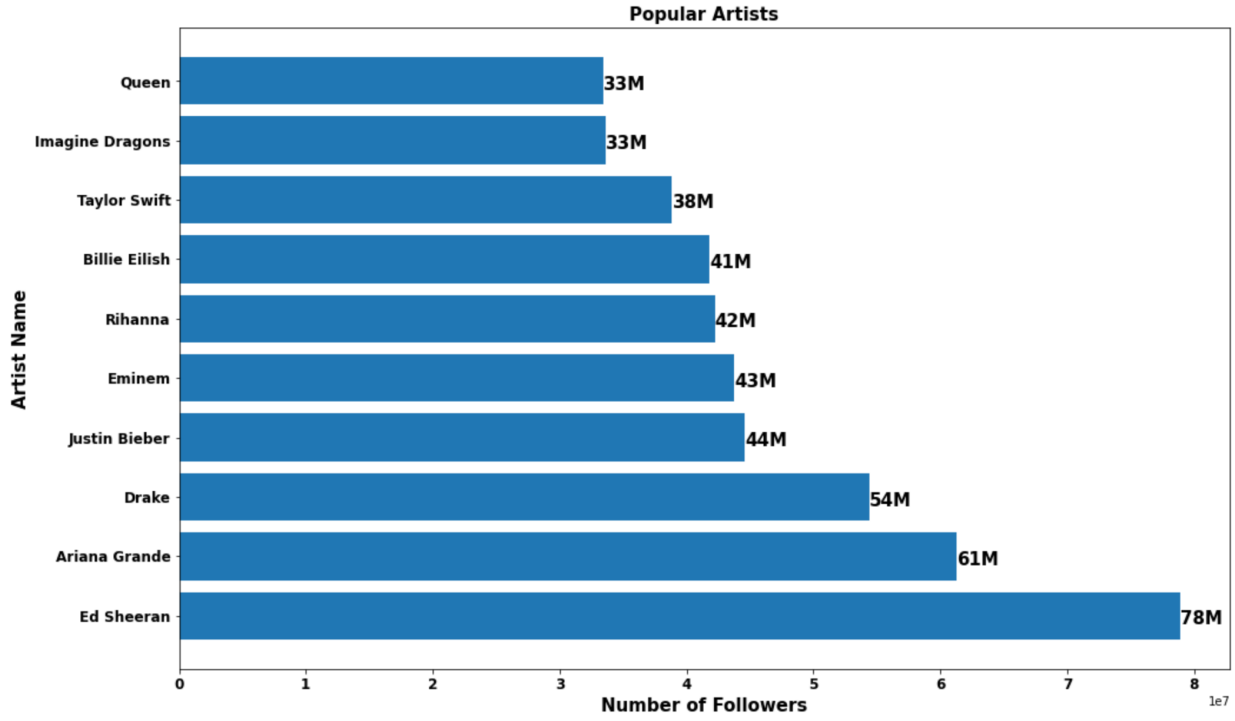
As we can see the above visualization shows the top 10 most popular songs, the most popular song name is “driver license” with 99 popularity measure.

As next visualization I studied the most popular artists based on the popularity measure.



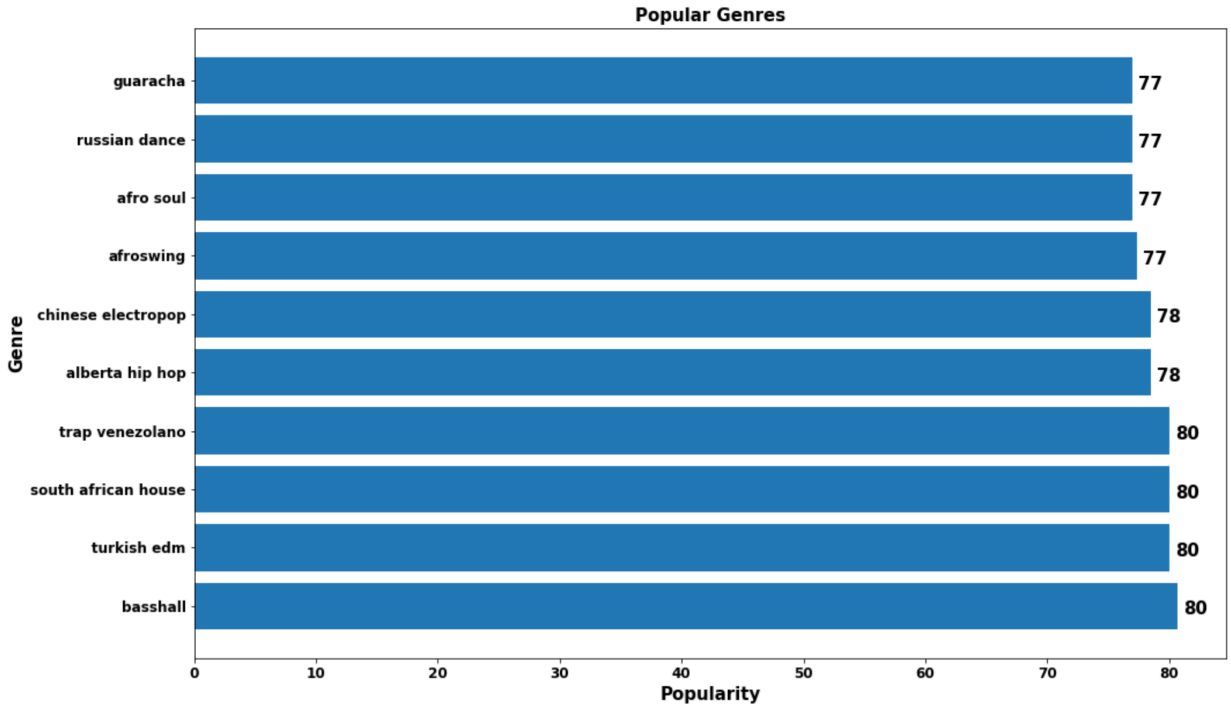
The above chart shows that Justin Bieber is the most popular artist with popularity measure of 100.

The next visualization is to know the popular artists based on the number of followers.



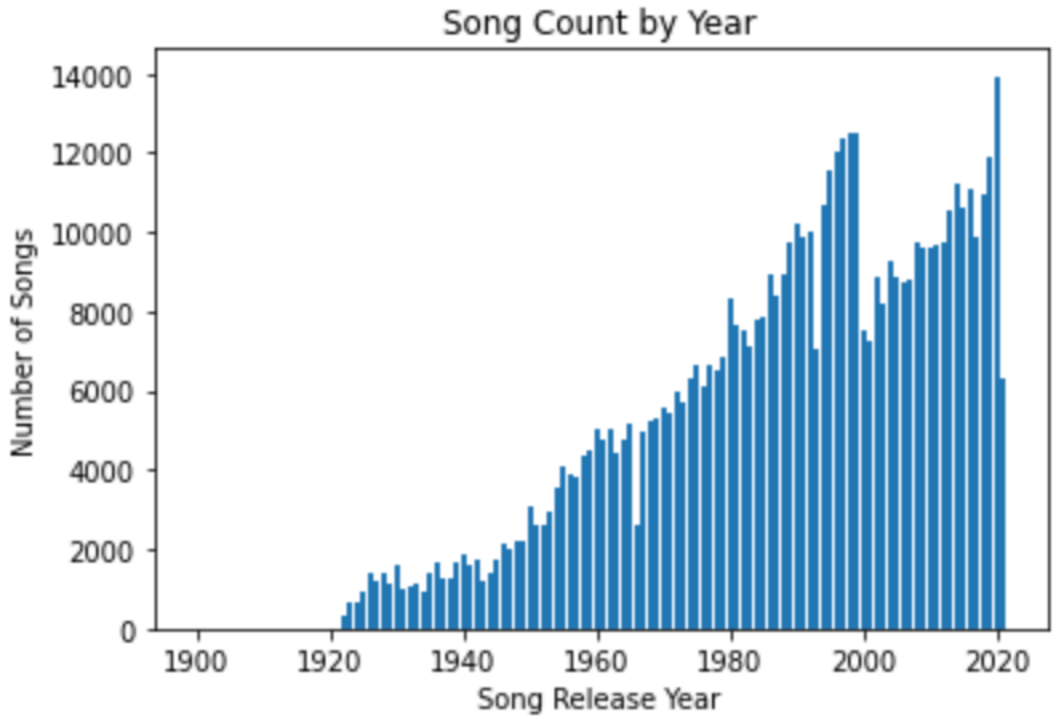
As we can clearly see that the “Ed Sheeran” has most followers. The above visualization shows top 10 artist with most followers.

It is also important to know about the genres, so I analyzed the topmost genres based on the popularity measure.



The above result shows the top 10 most popular genres. The genre with maximum popularity is basshall. Turkish edm, south African house and trap venezolano is also popular with same popularity measure 80.

I also analyzed the number of songs per year to understand how many songs release per year.

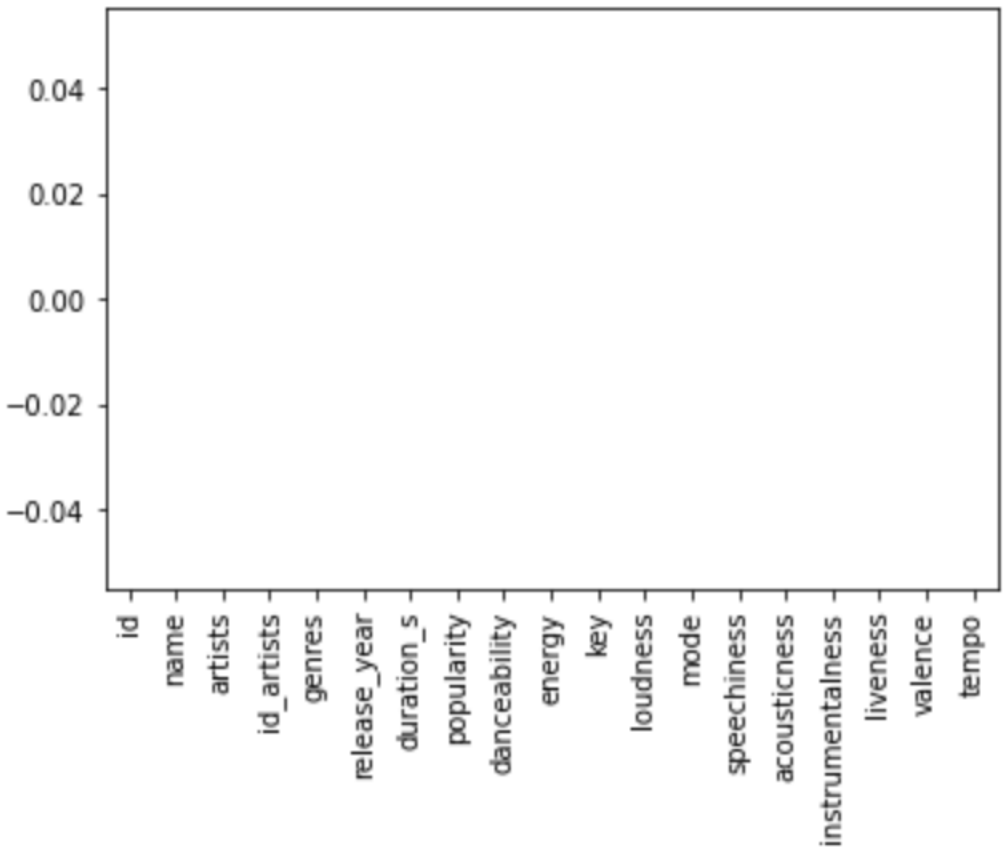


As we can see the count of the songs is constantly increasing per year, the most songs were created in year 2020 and in year 1990. This visualization shows the successful years in the music industry.

As next I performed some cleaning steps. Cleaning process is very important as without which we cannot use any data. The raw data has lot of noises that needs to be removed before we are building the model otherwise it can affect the performance of our model and can lead us to inaccurate results. I used isnull() function to see how many features have missing values in our dataset. The missing values can impact the result accuracy and reliability hence it is very important to handle them carefully in preprocessing step.

```
id                      0
name                   71
artists                71
id_artists              0
genres              49825
release_year            0
duration_s              0
popularity              0
danceability            0
energy                  0
key                     0
loudness                0
mode                    0
speechiness             0
acousticness            0
instrumentalness        0
liveness                0
valence                 0
tempo                   0
dtype: int64
```

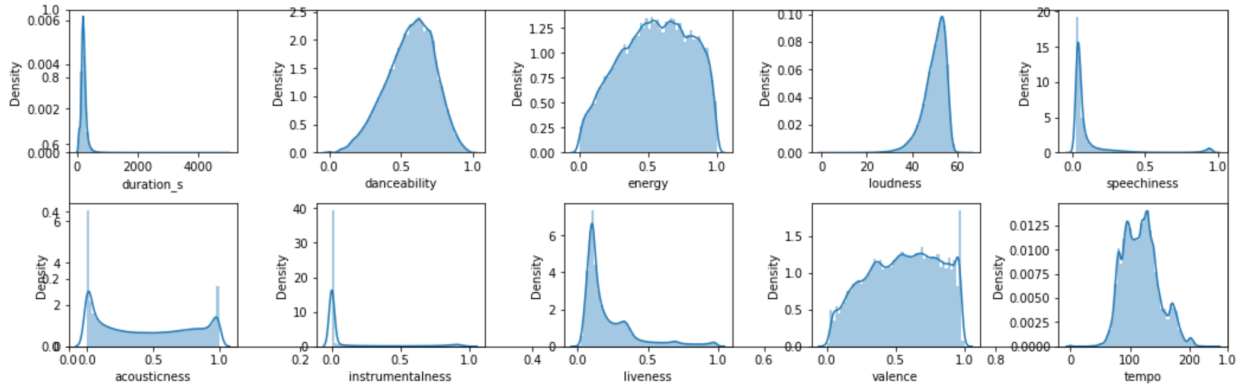
Looking at the above result we can clearly stat that column genres have 49825 rows with missing values. This is the most important feature, which will be used in song recommender system to recommend the songs to the users and hence it need to be treated. For resolving the missing values issues I am removing the 50,000 rows from our dataset. Our dataset has more than 500K rows and removing 50K rows will not impact the dataset. Let visualize the data again to see if there are any null values in the dataset after removing the data.



As we can see above there is no null values in the dataset.

We also noticed using the nunique() function that the dataset has some duplicate values. It is important to remove the duplicate values from the dataset before creating the system. In order to do that we drop the duplicates rows from the dataset using the drop_duplicate() function.

As next step I used distribution plot to understand the numerical features in the dataset.



As we can see above that some of the feature's distribution is skewed as well as we see some of the features are normally distributed.

Conclusion:

The system is built using the cosine similarity, this system will take the input as song name and compared the two vectors – the input song compares the similarities with existing dataset songs features and will recommend the top 5 songs. This model based on the cosine similarity model. The built system is tested and is working as it recommended the songs based on the input song.

```
recommend_songs('Elephant Love Medley')
```

```
This song is either not so popular or you    have entered invalid_name.
 Some songs you may like:

You Never Know
Pa' Que Retozen
The Thrill
Feels Like Summer
Short Skirt / Long Jacket
```

```
recommend_songs('As Long As You Love Me')
```

	name	artists
115460	Intentions (feat. Quavo)	Justin Bieber, Quavo
82236	Cooler Than Me (feat. Big Sean)	Mike Posner, Big Sean
90372	Plain Jane REMIX (feat. Nicki Minaj)	A$AP Ferg, Nicki Minaj
373664	Bon appétit	Katy Perry, Migos
83160	Baby	Justin Bieber, Ludacris

Assumptions:

All the analysis has been performed using the dataset available online assuming that the data is representative of the larger environment. The biggest assumption is that this dataset is collected without any bias since we are not sure how this data is collected and the original purpose of collecting the dataset. Any bias in dataset can lead us to bias model and to inaccurate recommendation of outcome.

Limitations:

One of the limitations is we using the historical dataset, we are not sure how old is this dataset. Also, this dataset has some features, but we are not sure if we are missing any important features that can changes the recommendation of outcome then we build the model with certain limitation. In order to deal with this limitation, we need to try this model on different dataset and a proper review should happen on this model by expertise to understand if we are missing anything in this model.

Challenges and Risks:

With all the assumption and limitation of the dataset the biggest challenge is to make sure the model we build is accurate. Some of the potential risk we can see depending on how the missing data, misbalancing and scaling of the data is handled in the preprocessing steps as all these factors can lead us to inaccurate prediction and bias in model.

Future Uses/Additional Applications:

The model we build here using machine learning can help business owners to scale their business as this recommender system will encourage users to make some more purchase. This model will also help the users to get familiar with some new songs that has same genres and other features of their past choices in songs. All this information will also help business owners to understand the demand and artist can work on it to become more popular among the people.

In order to advance our machine learning model, we can analyze more tracks data that is available to us ethically by states wise or county wise to understand the demand and work on it to scale the artist popularity and users experience.

Recommendations:

To make our system better we can add collaborative filtering in our content-based system. In existing system, we are trying to recommend songs based on the genres and other features that are similar to the user past choice songs genres. With collaborative filtering we can recommend the songs based on the other users with similar tastes

This method will allow us to provide more personalized recommendation to the users.

Implementation Plan:

The next step is to partner with different business owners, share the importance of the machine learning model with them, get access and collect more data so the model can be tested more on real time data.

Once the testing is done, we can implement the model live where model can be used by users and can recommend songs to them. Once the model has all the features, it can recommend songs to the users. Regular checks can be done to make sure of the consistency of the model and flow of the data.

Ethical Consequences:

I make sure all the ethical analysis guidelines will follow by me while preforming data analysis steps.

1. To ensure that legal and ethical ways are used to collect the data and make sure no personal information of users get misused or disclosed without the consent.
2. Make sure that all the algorithm and models that are used in the project do not show any biases and discrimination towards any group.
3. There should be transparency in the data analysis methods so that everyone understands the analysis easily.
4. Will make sure the security of data so there will no unauthorized access to the data.

5. Finally, will ensure there will no negative impacts on population because of the project result.

References:

1. Data Source : - https://www.kaggle.com/datasets/subho117/music-recommendation-system-using-machine-learning/data