Model based Song Popularity Prediction using Spotify Data

Shaashwat Jain
dept. Computer Science and Engineering
PES University
Bengaluru, India
shaashjain213@gmail.com

Srishti Sachan
dept. Computer Science and Engineering
PES University
Bengaluru, India
srishtisachan31@gmail.com

Deeksha D
dept. Computer Science and Engineering
PES University
Bengaluru, India
deekshad132@gmail.com

I. Introduction

As the technology is increasing, people are becoming more driven by streaming apps like Prime Video,Hulu, Netflix etc. More music apps are being used by people all over the world. In our analysis driven world, the focus mainly shifts to these platforms to deliver the content which would be appreciated by the youth and all generations alike. The need to focus these interests and their marketing analysis on songs and video-based content is now more important than ever. A research showed that USA recorded music revenues which grew from 5.6% to $5.7 billion in the first half of the year 2020. Streaming music apps like Spotify grew to 90% compared with 80% the prior year.[1]
Clearly, good prediction models for music popularity can be very beneficial considering the present stats.
The present scenario has further increased the usage of music streaming devices. The revenue in the Music Streaming platforms is projected to amount to US$16,395 billion towards the end of 2020.This is also due to the fact that many people stayed home amid the COVID-19 outbreak and hence the increased usage of over-the-top (OTT) platforms. [1] These devices are able to search alike music and come up with a playlist according to the likeness of the user because of their recommendation systems. They can do so because of the (user generated) big data and their digital song database.
Having a fundamental understanding of what makes a song popular has major applications to businesses that thrive on popular music, namely radio stations, record labels, and digital and physical music market places. As it has a huge impact on business, researchers and product developers are looking for good product success/failure prediction models.

Music, other than being used for entertainment purposes can be used by many for influencing mood. The music we hear can have a positive or negative impact on our health according to the track played. If we can further the application of our paper to modify behavior in people, it can have a major impact on people with health conditions, specially clearing all the bad thoughts anyone has, hence reducing deaths by suicide and depression.

Music is also thought to be a way to communicate with others; it can also work as a romantic ambiance, people like to listen soft music during their emotional states like love. Specific music induces specific feelings in listener. It has also been proven through a research that people enjoy more when they listen to familiar music or their favorite one. Music is increasingly being used to enhance well-being, reduce stress and distract patients from unpleasant symptoms. Keeping all this in mind, it is important for a recommendation system to recommend music according to the likeness of the user. [2]

Going through various research papers, we came across different technologies and methods that have shown some improvement and modifications in the present recommendation and prediction models. Our research builds on as an extension to these studies and the results generated may prove to be beneficial for the music industry.
For this research, we study the Spotify database API with 160K+ entries from the years 1921-2020. The question of what makes a song popular has been studied before with varying degrees of success. Every song has key characteristics including duration(in ms), artist, tempo, key, valence, loudness, chord, etc. Previous studies that considered lyrics to predict a songs popularity had limited success. Also there is varying interest of people in music among different generations. Youth alone has a variation of interest in music like hip-hop, metal music, EDMs, acoustics and many more. This makes the prediction

complex.We intend to justify this trend with our project and get other important insights. We are going to try to build a model which can help us understand the trends and genres of each generation and predict what could be a game changer for the field from the many inferences.

II. Literature Survey

In this literature Survey we focused mainly on three papers. The first paper [3] , focuses on the popularity field of the data set and tries to explain what each field for the given time frame explains about the popularity of the song. The research question is: "Is the attribute approach based on Spotify's audio features effective in explaining streaming popularity on Spotify?". This helps us gain insights on how to classify the data set songs on the basis on genre so prediction of popularity can be made easy. The results of the correlations showed that there were significant relationships. The found relationships however were generally weak. Next the paper was built on a regression model from a selection of attributes by a step-wise method. Its explanatory power (R2) came out to be around 20.2%, meaning that the model explained 20.2% of the variation in the popularity. Therefore, it concludes that the model is not as effective in explaining popularity on its own. The popularity of the songs can be highly dependent on the other attributes like liveliness, acousticness, which is something that may help us with our prediction. As different genres do not share the same popular attributes, there will be noise in the hit prediction model making for a lower R2 value and lower correlations . We can modify the data set to remove the noise to improve our model.

The subsequent paper which gave us a brief insight about our data set [4] , gives a Generalized Linear Mixed Model (GLMM) based model for the analysis. This research investigates the relationship between song data, audio features obtained from the Spotify database (e.g. key and tempo) and song popularity, measured by the number of streams that a song has on Spotify. One assumption made by the paper is that the popularity is based, in the most part, on the total number of track plays and taking into account how recent those plays are. Songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. The popularity rate is also heavily dependant on the fact that now, the population has increased which can be a huge factor in determining popularity of the song. The main claims that the reputed authors have made are, speechiness, instrumentalness and Live are the features that negatively affect the Popularity Index, while Energy, Valence and Duration of the song are the ones that positively affect it. This research contributes to further understanding in the field of new product success prediction.

There are many more fields in the data set which can point us in the right direction to predict and correlate different parts of a song to its popularity and how well it will be received. Also, it can give us an insight to what kinds of playlists will be preferred by a specific demographic.

The last paper featured in our research out of the many more prestigious ones [5] , treads on the lines of genre prediction using machine learning on the top ranking songs. The main focus here is on clustering genres based on audio features. One major assumption made in this paper is the songs which have more than one genre, the first listed genre was selected as the song's genre i.e. primary genre. The authors claim that there are genres with similar combination of audio features, hence the Principal Component Analysis (PCA) was implemented to covert audio features that are possibly correlated into linearly uncorrelated principal components.

- Sound component: energy, loudness, acousticness
- Word component: speechiness, valence
- Rhythm component: danceability, liveness, valence

One of the major takeaways from this paper was to make sure our features are uncorrelated by taking into account all its genres and also classify a song into multiple genres.

The highly extensive survey done above gives us an insight on what lines to follow while going through our analysis and also a very good insight on which field from our data set could point us in the right direction to get some meaningful results at the end.

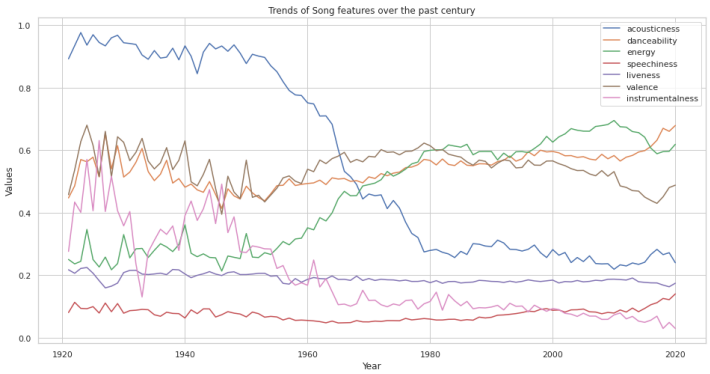
III. EDA & Visualization

Our initial dataset shape is (169909, 19). The number of duplicate or missing rows is 0.
We drop the column 'id' because it is just an identifier and we can gain no insights from it.
Since most of our features follow a normal distribution, we use z score with a threshold of 3 to remove the outliers from our dataset.

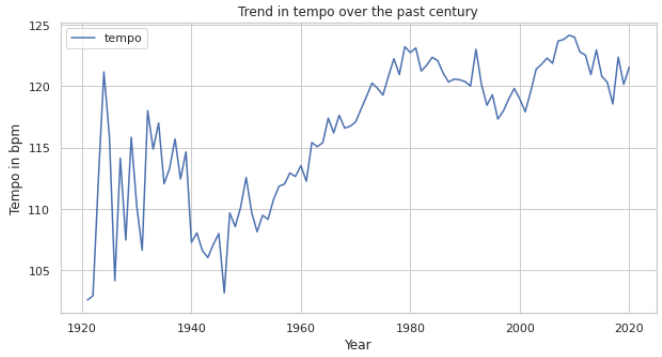
$$z = \frac{X - \mu}{\sigma}$$

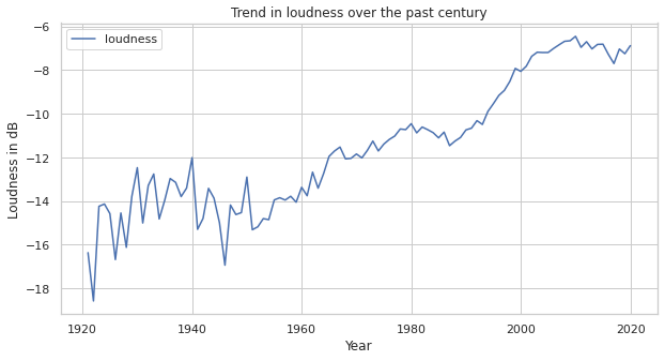
Fig. 1: X : observed value, μ: mean of data, σ : standard deviation of data

We get the cleaned dataset which is of shape (157218, 18)
We observe the high positive correlations between Year and Popularity (0.88) and Energy and loudness (0.79)
There also exists negative correlation between Energy and acousticness (0.77), Year and acousticness (-0.66) and Popularity and acousticness: -0.62

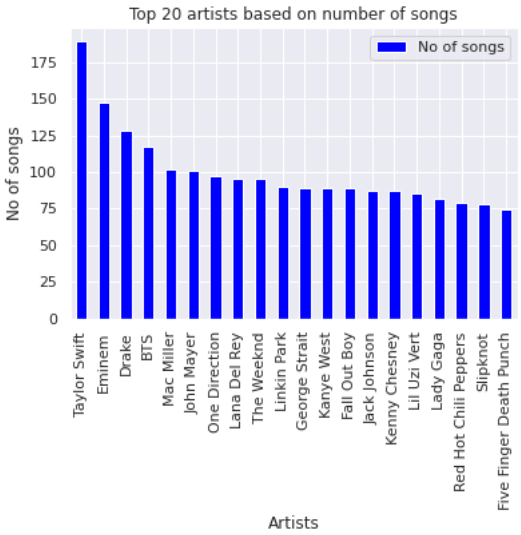


Over the past century, the acousticness and instrumentalness of songs have drastically decreased and songs trends are moving towards higher valence and energy.

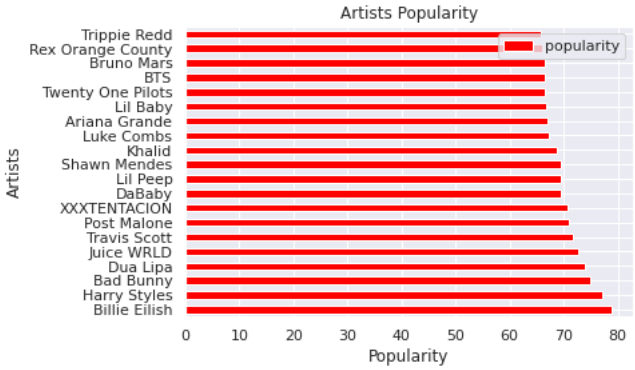




Both the tempo and loudness show a similar increasing trend over the years.
We sort the artists based on their number of songs. Here we see that Taylor Swift has the most number of songs. We will be analyzing later if more number of songs is directly proportional to the popularity of the artist.



We next calculated the mean popularity of all songs released by an artist.



This shows that artists who release more songs are not necessarily the ones with the highest mean popularity. Take for example Taylor Swift, who has the highest number of songs does not feature in the top 20 artists.
The songs released after 2000 show a 22.5% of songs labelled as explicit compared to the 4% of explicit songs released which is a significant increase and thus a strong correlation to the changing generation and trends.

IV. Model Building

To further our analysis on the preprocessed data we logically processed some of the strong yet powerful correlations for approaching an accurate prediction model.

Looking onto predicting our variable 'popularity' of the song we first had the relation between a specific time range i.e. generation which we abstracted to as a century and then we have a relationship to identify which artists can get higher ratings which will not depend on the artist who publishes more songs in their career.

The models chosen for implementation and testing are:

- Logistic Regression
- Support Vector Regression
- Decision Tree Regressor
- Multi-Layer Perceptron
- Random Forest Regressor

To further on this, the initial model approach was to use a Logistic Regression model so we can regress on the given variables of any song excluding the popularity and artist information. Data is fit onto a Linear Regression model, which can be acted on by a logistic function. Logistic regression outputs an estimated probability of what our popularity is going to be. Taking this estimate, we further our result and calculate a predicted value for probability which we can compare with our actual results and calculate an accuracy or error rate with.

The next supervised Learning model looked onto is Support Vector Machine(SVM) which has been used for regression analysis. SVM being a classification model can be extended to do regression analysis also known as Support Vector Regression(SVR). In our case the data is non-linear as we have shown at the preprocessing and analysis stage of our paper.

For non-linear data the SVR model uses a trick called the kernel trick, where it estimates the data points using a kernel estimator i.e. a mathematical function. There are multiple kernel functions : linear, polynomial, rbf and sigmoid. For our analysis we have used the sigmoid kernel.

The next step on building an accurate model for prediction was Decision Tree. Decision Tree being a generic classification model can also help us predict individual popularity metrics of each field in our dataset clubbing which we may get a regressed model with integer values for our popularity field. This model is also called the Decision Tree Regressor. Going over this approach we fit our entire dataset randomly sampled and train it to fit and calculate our error metrics.

One of the advanced models we looked into is the Multi-Layer Perceptron model which uses Neural Network techniques to calculate our popularity field. The model being a Neural Network model needed hyperparameter tuning, after which it gave us better results than expected. We used the 'Relu' activation function to process our data. Using multiple hidden layer over thousands of iterations we got our model to as high as a 70 - 80 % accuracy which is discussed in the next section also.

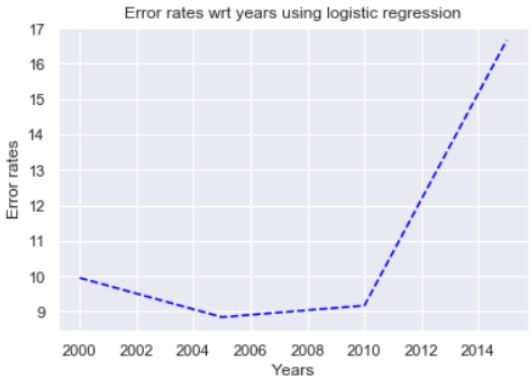
The final model we looked into is an extension of the Decision Tree Regressor we got earlier. As being a basic prediction out of the first three, Decision Trees got us a higher accuracy when compared. To use this to our benefit we used the technique called Random Forest Regressor which uses multiple Decision Tree Regressors to get more accurate results. After we tuned the hyperparameters for this model we got similar, if not better, results compared to the Multi-Layer Perceptron model.
After testing various models as such, we have presented the results and certain inferences in the below section.

V. Experimental Results

We have taken 5 models for our approach: Decision Tree, Logistic Regression ,Support Vector Regressor, Multi-Layer Perceptron Regressor and Random Forest Regressor.

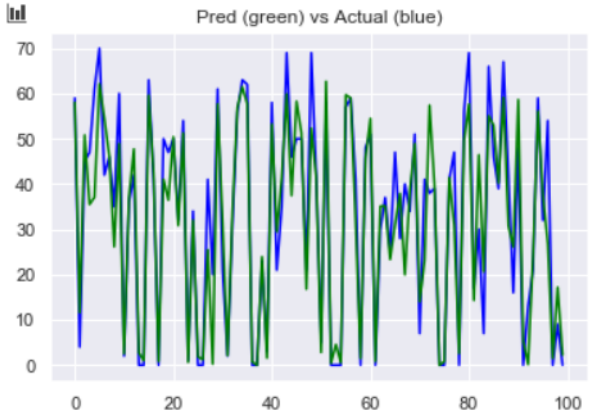
- Decision tree
  - With this model, we get a final error rate of approximately 14 for our predictions after the year 2015.
  - We have trained our model on the sample after the year 1999.
  - This was done to give us a better idea of the trend.
  - We calculated the error rates for every 5 years starting from the year 2000.

- Logistic Regression
  - With this model, we got an error rate of around 9.95 initially but then it started to increase with the years.
  - This can be due to the fact that as the trend in music changes, it is hard to predict popularity because there are so many features which are periodically changing with trend.
  - The last recorded error we got was 16.66 for this model.



- Support Vector Regressor
  - The model was again trained over random samples from the whole dataset.
  - Initially our error rate is 10.30 and finally comes out to be at an average 16

  - In the trend of the errors across the years, we see that it decreases till the year 2010 then starts increasing from 2010 to 2015.
- Multi-Layer Perceptron Regressor
  - Multi-layer perceptron with 3 hidden layers of size (8,4,6)
  - Activation function used for testing: "relu"
  - Obtained an r2 score of approximately 77%.



  - The graph above shows how close the predicted andd actual value are and when are the times they do not coincide
- Random Forest Regressor
  - We have used 10 splits of the data i.e. tree
  - The training was repeated for 3 times
  - A single Decision Tree Regressor has r2 score of 74% along with MAE of -6.495 and 0.647 as shown below

```
MAE: -6.495 (0.647)
The R2 Score is  0.7431393434682504
```

VI. Conclusion

Our analysis shows that audio features do play a vital role in the popularity of a song. A discussion on the trends of music features over the past century has been provided. On testing various models, it is seen that Decision Tree Regressors, SVM and logistic regression do not yield good results. Random forests and Multi-layer perceptron perform very well even for small-sized samples. With better optimizations and more complex models from deep learning such predictions can be used in the real world to help artists.

References

[1] "MID-YEAR 2020 RIAA REVENUE STATISTICS", 2020. Available : https://www.riaa.com/wp-content/uploads/2020/09/Mid-Year-2020-RIAA-Revenue-Statistics.pdf

[2] Ahmad, Nawaz & Rana, Afsheen. (2015). Impact of Music on Mood: Empirical Investigation. Research on Humanities and Social Sciences. 5. 98-101.

[3] Nijkamp, Rutger, "Prediction of product success", Explaining song popularity by audio features from Spotify data, 2018. Available : http://purl.utwente.nl/essays/75422

[4] Sciandra, Mariangela & Spera, Irene. (2020). A Model Based Approach to Spotify Data Analysis: A Beta GLMM. SSRN Electronic Journal. 10.2139/ssrn.3557124.

[5] Luo, Kehan. "Machine Learning Approach for Genre Prediction on Spotify Top Ranking Songs." (2018).

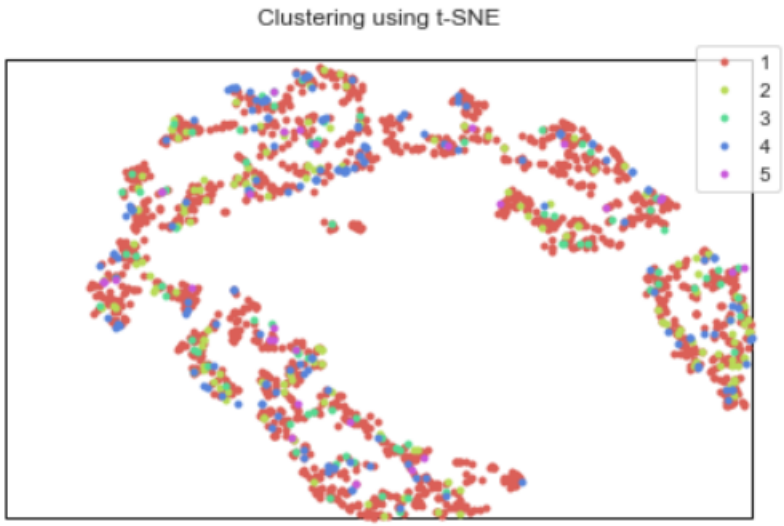
VII. Additional Information

Contributions:

- EDA : EDA and visualization was done equally among all team members
- Model Building: This part was divided among us all.
  - Decision Tree handled by Srishti Sachan
  - Random Forest and SVR was handled by Deeksha D
  - Multi-Layer Perceptron and Logistic Regression handled by Shaashwat Jain
- Results: All of us contributed with major contributions by Srishti Sachan

Additional Insights:

We used t-SNE for our dataset. It is a dimensionality reduction technique which works by converting similarity between data points to a joint probability. We used this technique to see if there is a hidden pattern which correlates different songs without actually providing the labels.



We found out songs which are more similar to each other but their was no hidden correlation.