Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

17

Songs Popularity Analysis Using Spotify Data: An exploratory study

Prathyusha Beesa, Vaishnavi Naregavi, Junaid Imandar, Surabhi Thatte

Department of Computer Science and Applications, School of Computer Science & Engineering,

Dr. Vishwanath Karad MIT World Peace University, Pune, India

Abstract:

This study presents an overview of analytical model for observing various factors which are impacting the songs popularity and predicting songs popularity using various machine learning algorithms. The data is collected using various methods. In most of the studies we found that researchers used Kaggle dataset and while others scrapped Spotify website to curate their own dataset.

We also found that maximum number of researchers predicted popularity of songs using same number of features of the songs i.e., Danceability, Tempo, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence.

We also, observed that all the researchers used unsupervised and supervised machine learning algorithms to prognosticate songs popularity. In future, researchers can investigate the use of deep learning and other neural networks to observe the performance. We also recommend that choice of appropriate data features and loss functions can ensure optimized outcomes.

We also aim to analyse the preferences of the songs by users before and after covid pandemic.

Key Words: Spotify, Music, Audio Features, Supervised Machine Learning, Unsupervised Machine Learning.

Introduction:

Preliminarily we hear songs by Radio or Television. But in Ultra-Modern Days we can listen to our favourite music just by downloading music application in our smart phones. Various music applications like Spotify, Gaana, Jio Savan, Prime Music and more available in the

market. But Spotify is famous among music streaming services users and popular music application in the market.

Spotify is the most admired online service that allow its subscribers to access to various digital contents like music, podcast etc. This service allows users to hear any content with minimal hardware requirements. It also allows users to create libraries of your choice and share among your favourite people.

The best thing about Spotify Platform is providing talented artists across the world to upload their content without the requirements of the record label and introducing interactive interface to build meaningful connections among users and artists.

Spotify provides music from various genres and artists: from Indie Rock to Top 40 Pop, movie soundtracks and classical music, spiritual songs, and podcasts of multiple categories. Spotify holds greater than 450 million users and approximately 195 million subscriptions across the world. At certain point of time, we all thought among various music applications why Spotify is so popular?

The answer is veritably simple to this question the recommendation system and search capability of Spotify. Spotify uses machine learning algorithm to analyses the preferences and historical data of the users and recommend the songs. It uses various machine learning algorithms and deep learning to continuously ameliorate its search algorithm. Now, let's understand the terms data science and machine learning and how these concepts are helpful to analyse and drawn meaningful insights from data.

[1] Data Science is quickly developing niche area which is considered to be an intersection of mathematics, statistics, programming, analytics, and artificial intelligence; used to uncover the insights from the big data. These insights are used in decision making by many organizations, where in they rely on data science to provide them with accurate predictions or recommendations.

According to latest information everyday approximately 328.77 million terabytes of data is generated. To analyse this huge amount of big data, Data science, Machine Learning, Artificial Intelligence and Deep Learning concepts are used. This the main reason why Data Science is one of the fast-growing fields in today industry. The popular data science tools i.e., python, R

etc, which can help in drawing insights from data.

Data Science is used in many sectors like Health Care to build health instruments to detect and cure diseases and used by many logistics companies to find faster routers for delivering services. Apart from these, it also used in Gaming, Retail Sector, Banking Sector and many more.

[2] [3] Machine Learning is a subset of Artificial Intelligence, which gives information about the software applications which tries to predict as accurate as possible outcomes of various models' outcomes without being explicitly programmed. It mostly uses the historical data as the input which can predict the future output. Machine Learning is broadly divided into two types Supervised Machine Learning and Unsupervised Machine Learning.

Supervised Machine Learning algorithm uses the labelled data. The model is trained on both input and output data. When the new data is given, it needs to identify by using the historical data. Some of the algorithms used are Binary Classification, Regression modelling, Ensemble Learning.

Unsupervised Machine Learning Algorithm uses the unlabelled data. It identifies the patterns among the data to predict the output. It uses clustering algorithms to identify the trends of the data. The algorithms used in unsupervised learning are Clustering, Anomaly detection, Association mining, Dimensionality Reduction.

Machine Learning is used in various sectors. Let's discuss few of them.

1. In Agriculture, the devices are built using machine learning algorithms which can detect the disease on plants and provide the solutions. It can also detect the nutrition level of the soil and can give information of which crops can be grown.
2. Twitter uses machine learning algorithms to detect the regionalist tweets and take action accordingly.
3. Machine learning algorithms used in crime department for the Facial recognition of criminals or terrorists.

The goal of this review paper is to investigate the choice of music users like to listen and analyse the songs popularity based on the various features impacting. At the same time

analysing the machine learning models used to predict songs popularity using Spotify Data.

Literature Survey

To understand the use of ML in this domain we reviewed the latest papers based on the type of data they used, models which were implemented to predict the song popularity, what results they obtained, how much accuracy they received and what are their drawbacks and advantages to understand the future scope of this domain for further research. The table below list the same in a systematic way for better understanding.

Table1: Literature review table.

GLMM: Generalized Linear Model with Mixed Effects, AIC: Akaike Information Criterion, CAIC: Conditional AIC, RF: Random Forest, LR: Linear Regression, SGD: Stochastic Gradient Descent, GBM: Gradient Boosting Model, SVM: Support Vector Machine, DT: Decision Tree.

Title	Models Implemented	Results	Accuracy	Future Work
What makes a song trend? Cluster analysis of Musical Attributes for Spotify top trending songs.	K-means Algorithm, Silhouette method, Agglomerative clustering	After analysing the data researchers noticed high correlation between loudness and energy whereas low correlation between valence and loudness	Observed the popularity of song increases if the danceability is High and instrumentalness is low	if the model is optimized and the analysis is done large sample data will lead to more deep analysis.
SpotiPred:	K means Algorithm LRRF	Using audio Features and genres of top-	RF- 95.37%	If more audio features are included in the sample data can improve the accuracy
Machine Learning Approach		ranking songs in the Spotify a		

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

Prediction of Spotify Music popularity by Audio Features		prediction model is created which can predict the top-ranking songs.		
Spotify: You Have A HIT!	K-Nearest Neighbour Classification with Down Sampling Gradient Boosting Model RFDT Logistic Regression SGD	By implementing the model using LASSO scaled linear correlation and it also give inputs that using underlying song features can predict the song can be commercial hit or not.	K-Nearest Neighbour Classification with Down sampling – 70.2% GBM- 71.83 RF – 63.89% DT – 67.72% Logistic Regression – 72.49% SGD – 72.51%	Only a thesis is made we can look for some research in the future by implementing this model.
A model-based approach to Spotify data analysis: A Beta GLMM	Beta Regression GLMM	It basically uses Beta GLMM to explore data catching section by Spotify Web API and propose statistical models for data analysis	AIC and CAIC are used for evaluating the model	This research will help in the contribution of the field HSS.
The music industry in the streaming age: Predicting the success of a song on	XGB Classifier LGBM Classifier RF Classifier Gradient Boosting Classifier	The prediction model is built to predict song success and draw insights along	XGB Classifier- 87.45% LGBM Classifier- 87.35% RF Classifier- 87.44% GB Classifier-	If the artist information is also included in the model can improve the accuracy and can get more analyses.

Spotify.	AdaBoost Classifier DT Classifier	with the features of the song's artist information is also equal	87.34% AdaBoost Classifier- 85.35%	
	Logistic Regression	important to anticipate the song success	DT Classifier- 82.53% Logistic Regression- 62.83%	
Predicting the Song Popularity using Machine Learning Algorithm	LR Polynomial Regression Lasso Regression DT Regression SVMDT Classification Perceptron Ensemble learning Voting Classifiers RF AdaBoost Gradient Boosting Bayesian optimization in hyperparameter Tuning	Here the model is built considering only audio signal related data and it was predicting the unpopular songs then popular songs due to imbalance data.	Linear Regression- 82.92% Polynomial Regression- 84.21% Lasso Regression- 83.19% DT Regression- 85.65% SVM -91.2% DT Classification- 88.68% Perceptron-81.33 Ensemble learning-92.11 Voting Classifiers-92.11 Random Forest- 89.54 AdaBoost-88.20 Gradient Boosting-89.11	In future research we can consider audio signals combined with audio features to anticipate songs popularity
Music intelligence: Granular data and prediction of top ten hit	Logistic Regression is implemented on three	The audio features are Divided into main and auxiliary	65% 67% 68%	As they considered only acoustic features. If PCA is used for feature selection can improve the accuracy.

songs.	models. Model 1: Only independent variables Model 2: Only main acoustic audio features Model 3: combination of main and auxiliary acoustic audio features.	acoustic features and the model is built using this data. It analysed that granular data provided by music intelligence technologies can help to make better decisions		
Popularity Prediction of music based on Factor Extraction and Model Blending	DT RF KNN algorithms	Using Linear blending of mentioned algorithms the researchers analysed that valence, speechiness and beats per min are most correlated and using PCA can give better results	MSE is used to evaluate the model DT-35.757 RF–18.808 KNN – 18.599 With linear blending these above algorithms the MSE came down to 4.96	Using large sample data can give better results.
Song HIT prediction: Predicting Billboard Hits using Spotify Data	Logistic Regression Neural Network RFSVM	The model is tested both on validation data and test data. It gave insights that audio features combined with artist past information can give variance in	Logistic Regression-80.65% and 81.51% Neural Network-82.14% and 83.05% RF-88.7% and 87.7% SVM-82.8% and 83.9%	PCA can be used to achieve better results and this model is used for artists and vendors to know which songs can be a HIT.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

		the data.		
Music Popularity Prediction through Data Analysis of Music's Characteristics.	LR KNN RF	It gave insights from heatmap that valence and BPM of song features are important for the song to rank high	Root Mean Square LR– 3.12 KNN – 3.3 RF – 4.5	It can be used by many artists and music vendors before releasing their songs and the model only built using text data instead of audio signals
Prediction o f Product success: Explaining song popularity by audio Features from Spotify data.	LR is built with SPSS	It analysed that along with audio features, artist information, Spotify stream count is also Equal important to anticipate songs popularity	Explanatory power (R2) is 20.2%	The model can build using other prediction algorithms like Decision Tree, Support Vector Machine, Random Forest etc to see the results.
A model for predicting Music Popularity on Spotify	SVM Gaussian Naïve Bayes Algorithm LR KNN	The dataset considered to build this model is Spotify Top 50 Ranking songs and Viral 50 public playlists and trained it using audio features.	SVM-90.81 Gaussian Naïve Bayes Algorithm- 84.56 LR-82.35 KNN-87.13	For future work, along with this data it can be also combined with artist popularity data to analyse it better.
Predicting Music Popularity Using Music Charts	AdaBoost Bernoulli Naïve Bayes Gaussian Naïve Bayes RF	The model is built to predict whether the Songs will represent in the Spotify's	AdaBoost 88.28,88.67 Bernoulli Naïve Bayes-88.49,88.69 Gaussian Naïve	To improve accuracy of the model, consider large sample dataset.

	SVM Linear SVM Poly SVM RBF SVM Sigmoid	Top 50 Songs.	Bayes-83.59,82.86 RF-87.30,88.47 SVM Linear-88.49,88.49 SVM Poly-88.49,89.01 SVM RBF-88.49,89.09 SVM Sigmoid-78.56,78.87	

Discussion

1. **Data:**

By analyzing all the papers, we can observe that most of them considered audio features like Danceability, Tempo, Valence, Acousticness, Instrumentalness etc., in common to anticipate songs popularity. One of the research papers [4] divided the song features into main acoustic and auxiliary acoustic features. Rest of the papers considered even artist information, stream count of the songs to anticipate songs popularity. Since all of them are using the same features, it is challenging in terms of feature selection and performing data pre-processing as there is not much scope with respect to converting the raw data into a quality data for model training.

2. **Model:**

Most of the Research papers implemented Machine Learning algorithms like LR, RF, SVM, DT, Logistic Regression, KNN. Some the models executed feature selection i.e., PCA to improve the accuracy.

LR: Supervised machine learning algorithm. It used to determine the relationship between dependent and independent variables.

DT: It is supervised machine learning algorithm. It consists of root node, branches, internal and leaf nodes. The internal nodes are outcoming branches of the root node which does all the possible calculations and then send it to the leaf nodes which consists of all possible outcomes.

RF: Supervised machine learning algorithm. Random Forest uses multiple decision trees to achieve the result. It used for both classification as well as regression cases.

SVM: It is also a Supervised machine learning algorithm. It handles both regression and classification problems. But in most cases, it is used for classification problem. SVM is to find a hyperplane in an N-dimensional space which classifies the data points.

Logistic Regression: It is a supervised machine learning algorithm. It is mainly used for classification problems. The main goal of the logistic regression is that to predict whether the data point belongs to the particular class or not.

KNN: KNN is a supervised machine learning algorithm. A typical classification problem which classifies the data point based on the characteristics. It has the assumption that similar data point can be found beside one another.

PCA: Is a dimensionality reduction technique. It is an feature selection technique to extract important features from huge amount of data.

3. **Outcome:**

In all research papers we can observe that the researchers used supervised machine learning algorithms to anticipate songs popularity considering song features.

In few papers they predicted whether songs will appear in the Top 50 songs or not based on the Billboard songs data.

We can also conclude that in this case RF the most robust model which gives accuracy greater than 80% and Logistic Regression model is not performing well which has accuracy less than 70% in most of the cases.

4. **Accuracy**

Accuracy is an evaluation metric which evaluates model performance of the classification problems.

$$\text{Accuracy} = \frac{number\ of\ predictions}{Total\ Number\ of\ Predictions}$$

Limitations

Through this review we found the following major limitation in the work done so far. They may be found useful for deciding the future course of research in this domain.

1. Most of the research papers considered supervised machine learning algorithms for predicting popularity of songs.
2. The dataset is also taken from Kaggle, and only limited data is allowed to be scrapped from the Spotify website. This makes all these models data agnostic.
3. All the models from the above papers are evaluated using accuracy metric. However, for classification problems there are other important metrics like sensitivity, specificity etc. which can be considered to improve model's overall performance.
4. After going through all papers, journals and articles I can say there is no single paper studied about the pattern of songs popularity before and after covid 19 pandemic.
5. All the researchers almost considered same song features to anticipate songs popularity.

Conclusion

At last, we can conclude that a lot more research can be done in this particular domain. We need to find ways to collect large sample dataset with various song features. Instead of only using supervised machine learning algorithms we can try unsupervised machine learning and deep learning algorithms to see results. Also, instead of using only accuracy as an evaluation metric for the model performance we can also consider precision, recall.

We can also think to implement loss cost functions, hyper parameter tuning to observe the impact on the data and results.

In future, we can also analyse how did the user preferences in the songs changed before and after covid pandemic.

References

[1] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, 2013.

[2] G. learning, "What is Machine Learning? Defination, Types, Applications, and more," Great Learning, 7 Feb 2023. [Online]. Available: https://www.mygreatlearning.com-/blog/what-is-machine-learning/. [Accessed 15 April 2023].

[3] E. Burns, "Machine learning," TechTarget, [Online]. Available: https://www.techtarget.-com/searchenterpriseai/definition/machine-learning-ML. [Accessed 15 April 2023].

[4] S. T. Kim and J. H. Oh, "Music intelligence: Granular data and prediction of top ten hit songs," *Decision Support Systems*, 2021.

[5] Z. Al-Beitawi, M. Salehan and S. Zhang, "What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs," *North American Business Press*, vol. 14, pp. 79-91, 2020.

[6] J. S. Gulmatico, J. A. B. Susa and M. A. F. Malbog, "SpotiPred: A machine learning approach prediction of Spotify music popularity by audio features," in *IEEE*, 2022.

[7] C. E. Dawson Jr, S. Mann, E. Roske and G. Vasseur, "Spotify: You have a Hit!" *SMU Data Science Review*, vol. 9, p. 5, 2021.

[8] M. Sciandra and I. C. Spera, "A model-based approach to Spotify data analysis: a Beta GLMM," *Journal of Applied Statistics*, 2022.

[9] M. Matera, "The Music Industry in the Streaming Age: Predicting the Success of a Song on Spotify," Universidade NOVA de Lisboa (Portugal), Portugal, 2021.

[10] Y. Essa, A. Usman, T. Garg and M. K. Singh, "Predicting the Song Popularity Using Machine Learning Algorithm," 2022.

[11] Y. Ge, J. Wu and Y. Sun, "Popularity prediction of music based on factor extraction and model blending," in *IEEE*, 2020.

[12] K. Middlebrook and K. Sheik, "Song hit prediction: Predicting billboard hits using spotify data," *arXiv preprint arXiv:1908.08609*, 2019.

[13] J. Kim, "Music Popularity Prediction Through Data Analysis of Music's Characteristics," *Int. J. Sci., Technol. Soc.*

[14] R. Nijkamp, "Prediction of product success: explaining song popularity by audio features from Spotify data," University of Twente, 2018.

[15] C. V. S. Araujo, "A Model for Predicting Music Popularity on Spotify," *Recall,* 2020.

[16] M. A. P. Araujo and R. Giusti, "Predicting music popularity using music charts," in *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA) IEEE*, 2019.

[17] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *American Association for the Advancement of Science,* 2015.

[18] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 Fourth international conference on computing communication control and automation (ICCUBEA) IEEE*, 2018.

[19] IBM, "Data Science," IBM, [Online]. Available: https://www.ibm.com/in-en/topics/data-science. [Accessed 15 april 2023].

[20] Simplilearn, "What is Data Science: Lifecycle, Applications, Prerequisites and Tools," Simplilearn, 09 March 2023. [Online]. Available: https://www.simplilearn.com/-tutorials/data-science-tutorial/what-is-data-science. [Accessed 15 April 2023].