International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

Spotify Stream Prediction using Regression Models

Aakash Kapadia[1], Tanvi Khasgiwal[2], Leena Kirtikar[3], Sanjay Pandey[4]

[1,2,3]*Student, Information Technology Department, Thadomal Shahani Engineering College, Mumbai*
[4] *Asst. Professor, Dept. Information Technology, Thadomal Shahani Engineering College, Maharashtra, India*

Abstract - *One of the biggest music platforms in the world is Spotify. With over a 100 million users per day and over 80 million tracks on it, people have been using this music streaming app for multiple things like podcasts, motivational audio and most importantly songs. Spotify is mainly used for songs as most of the worlds renowned artists to publish their content so the world can hear it and enjoy it. Being avid users of Spotify ourselves, we tried to find out what drives the fame of a song – or even try to understand why people listen to a specific song. In this research paper and project, we used multiple algorithms to check what attributes of a song makes it famous and why does it top the weekly charts. This paper uses a Spotify Database and performs Exploratory Data analysis on it to recognize the most influential variables and then further work on them. We have used algorithms like Linear Regression, Random Forest, Ridge Regression and Lasso Regression to compare accuracies of our admired results. Finally, the accuracy will be compared so that we can calculate the approximate streams of a song based on the relevant attributes.*

Key Words: **Spotify, streams, linear regression, ridge regression, random forest, lasso regression**.

1. INTRODUCTION

We chose a Spotify as our subject because for years as students, we have been using this app to help in various ways. Spotify aides us in entertainment with its millions of songs and several genres. It has many more advantages which makes it a better music app than other apps that exist. Firstly, it remains of the easiest apps to use and thus multiple age dynamics can be seen as users of the app. Spotify gives a great music sharing experience as friends and family can share songs using a shared account. The app also provides us with a premium version with no advertisements which enhances the experience by saving the music offline and not using network data to stream it. The biggest advantage is compatibility. It can be used over many devices including iOS, Android and Windows.

Finally, the music collection is what drives people on the app and that's why mainly the world uses it. We wanted to know why is a song famous, does the artist on the song matter, and what attributes of the song can affect the number of times it has been streamed. To answer our questions, we have implemented the following in our research work:

i) Using data visualization plots like bar chart, scatter plot, and heatmap we find the attributes which affect the streams of the song.

ii) We then move forward by dropping the attributes like Artist name, Song name, and its release date as in step 1 we realize that they do not affect the streams of a song by a larger scale.

iii) Characteristics of a song like Loudness, Popularity of the song, and its energy were the most contributing factors.

iv) Lastly, we take the regression algorithms mentioned and predict the streams of the song using the relevant factors.

2. LITERATURE REVIEW

Authors of research paper [2] investigated the connection between song information, such as key & tempo from the database of Spotify audio properties, and song popularity as determined by the numerous streams of Spotify. They researched four ML algorithms: Linear Regression (LR), Random Forest (RF), K-means, Clustering & created a highly accurate model for predicting success of particular songs. Their research offers a prediction model for figuring out whether a song will be well-liked by the general public and uses machine learning to categorize songs according to how well-liked they are.

In [3], Charts Carlos Vicente S. Araujo Marco A. P. Cristo Rafael Guisti make predictions on whether an existing well-liked music will garner greater than normal public interest and go "viral." They also make predictions about whether unexpected jumps in popularity will last over time. They base their conclusions on information from the streaming service Spotify, using "Most Popular" list as a proxy for popularity and its "Virals" list as a proxy for interest increase. Additionally, they take a classification approach to the issue and use a SVM model predicated on famous data to forecast interest and vice versa. Finally, they check to see if acoustic data can be helpful features for both tasks.

In [4], Elena Georgieva, Marcella Suta, and Nicholas Burton attempted to foretell which songs will top the Billboard Hot 100. They collected dataset of about four thousand popular & nonpopular songs and used web API of Spotify to extract the audio properties of each song. Using five machine learning algorithms, they were able to predict a song's billboard success with about 75% accuracy on the validation

set. The two most effective methods were neural network logistic regression.

In [5], Rutger Nijkamp's, study looks into the connection between music information, such as the key and tempo of a song & popularity of song as determined on the basis of number of song's streams received on Spotify. The attribute-approach was utilized to investigate the potential explanatory power of song qualities on stream count. 1000 tracks from ten different genres were examined via the Spotify database API. Regression was used to create a prediction model. With this research design, the results indicate that Spotify's audio features have minimal to average explanation power of greater count of stream.

3. METHODOLOGY

Methodology section of the paper explores the different methods which we have applied on the Spotify dataset to predict the number of streams of a song. The dataset is used in multiple cases of data visualization and predicting an approximate number. Fig 1 displays a flowchart of the methodology which has been used and implemented in our research and includes data cleaning, its display on graphs and usage of various regression models and obtain appropriate results.

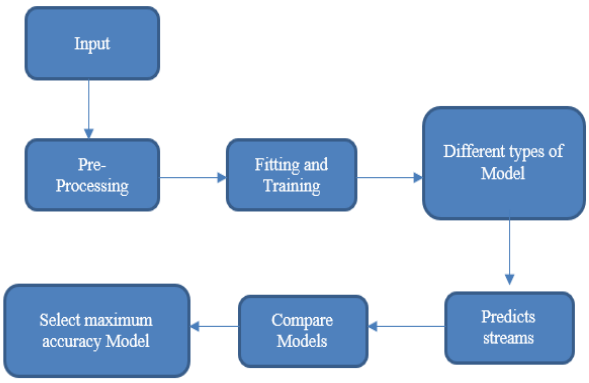

Fig. 1. Flow diagram of research work

3.1 Data Collection

The data is obtained from [1]. The dataset includes features like song name, artist, artist followers, genre, popularity, danceability, energy, loudness and many more.

Song Name	Streams	Artist	Artist Followers	Genre	Release Date	Popularity	Danceability	Energy	Loudness	Speechiness	Acousticness	Liveness	Tempo
Beggin'	48633449	Måneskin	3377762	['indie rock italiano', 'italian pop']	08-12-2017	100	0.714	0.800	-4.808	0.0504	0.1270	0.3590	134.002
STAY (with Justin Bieber)	47248719	The Kid LAROI	2230022	['australian hip hop']	09-07-2021	99	0.591	0.764	-5.484	0.0483	0.0383	0.1030	169.928
good 4 u	40162559	Olivia Rodrigo	6266514	['pop']	21-05-2021	99	0.563	0.664	-5.044	0.1540	0.3350	0.0849	166.928
Bad Habits	37799456	Ed Sheeran	83293380	['pop', 'uk pop']	25-06-2021	98	0.808	0.897	-3.712	0.0348	0.0469	0.3640	126.026
INDUSTRY BABY (feat. Jack Harlow)	33948454	Lil Nas X	5473565	['lgbtq+ hip hop', 'pop rap']	23-07-2021	96	0.736	0.704	-7.409	0.0615	0.0203	0.0501	149.995

Fig. 2. Dataset

3.2 Data Preprocessing

In this step we remove the inaccurate or null values directly from the .csv file of the dataset which might lead to wrong results. We had to perform less cleaning as the obtained dataset was nearly obsolete.

3.3 Data Visualization

Figure 6 shows a heat map of the data which we have used to achieve our results. Heatmap map is defined as graphical notation of large volume of data coded by different colours. The heat map takes all the variables present and correlates them with each other at the same time. Figure 7 shows a histogram of number of times a song was charted vs Song Name, figure 8 is a bar chart plotting of an artist vs streams and figure 9 is bar chart reflection of genre and streams.

3.4 Data Analysis

On further evaluation of the heatmap, we came to a conclusion that the most prominent features are Loudness and Energy. This can also be verified with the graphs shown in Figure 3 and 4.

Figure 3 shows us a scatter plot between popularity and loudness. The graph displays the relation that higher the loudness, higher the popularity of the song.

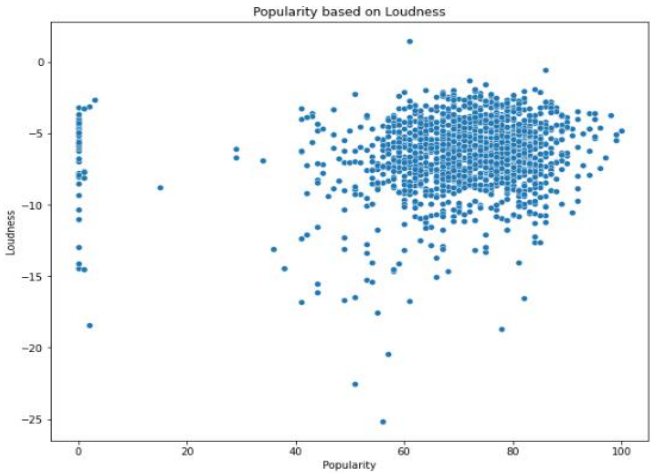

Fig. 3. Popularity Vs Loudness

Figure 4 shows a scatter plot between popularity and energy and it shows that higher the energy of the song, more the popularity.

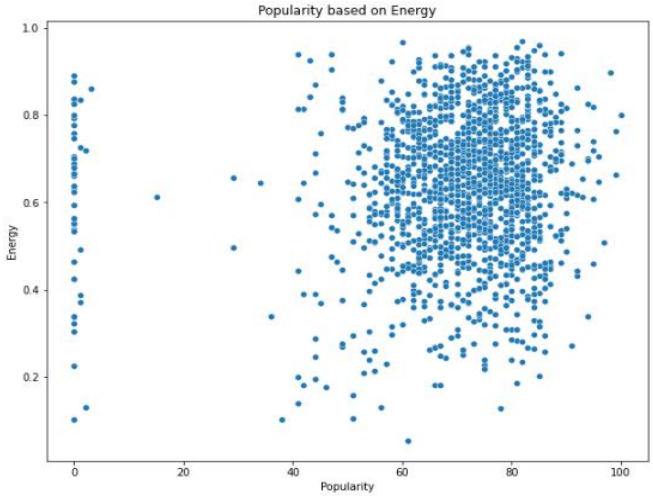

Fig. 4. Popularity Vs Energy

Figure 5 displays the regression models we have used with their respective accuracies. The regression model with the most accuracy is random forest.

Regression Models	Accuracy
Lasso	0.162616
Ridge	0.162591
Random Forest	0.974894
Linear	0.162616

Fig. 5. Accuracy Table

Linear Regression is a supervised machine learning algorithm which predicts the target value on the basis of the independent variables provided.

According to the accuracies found, we deduced that Linear Regression algorithm would provide us with the best possible results for our dataset.

The Hypothesis function of linear regression is stated below:

$y = \alpha 1 + \alpha 2.x$ ... (A)

here x denotes input training

y = labels to data

α1 = intercept

α2 = coefficient of x

Regression in Random Forest makes use of ensemble learning method. It operates by construction of many decision trees during the training period and provides us the mean of the classes as a prediction of all trees.

Here LASSO is used acronym of least absolute shrinkages selection operator. Extension of it is penalty term related to cost function. Summation of the coefficients is represented by this phrase. This particular term confine, forcing the model to curtail the value of coefficients in order to minimize the losses, when the value of coefficients increases from 0 to 1. In contrast to lasso regression, which commonly makes the coefficient absolute zero, ridge regression never puts coefficient value as zeros.

The LASSO regression combines statistics and ML to enhance predictability as well as understandability of the resultant model.

$$L_{lasso} = argmin_{\hat{\beta}} \left(\|Y - \beta * X\|^2 + \lambda * \|\beta\|_1 \right)$$

In essence, a regularised linear regressor is what a ridge regressor is. In other words, we add a regularized term to the initial cost function of the linear regressor in order to drive the learning algorithm to suit the data and help maintain the weights as low as feasible. The 'alpha' parameter of the regularised term controls the regularisation of model, therefore reducing the variance of the estimations. In some conditions where variables are independent & correlated, technique of ridge regression is used for forecasting the coefficients of numerous regressions.

Cost Function for Ridge Regressor:

$$J(\Theta) = \frac{1}{m}(X\Theta - Y)^2 + \alpha\frac{1}{2}(\Theta)^2$$

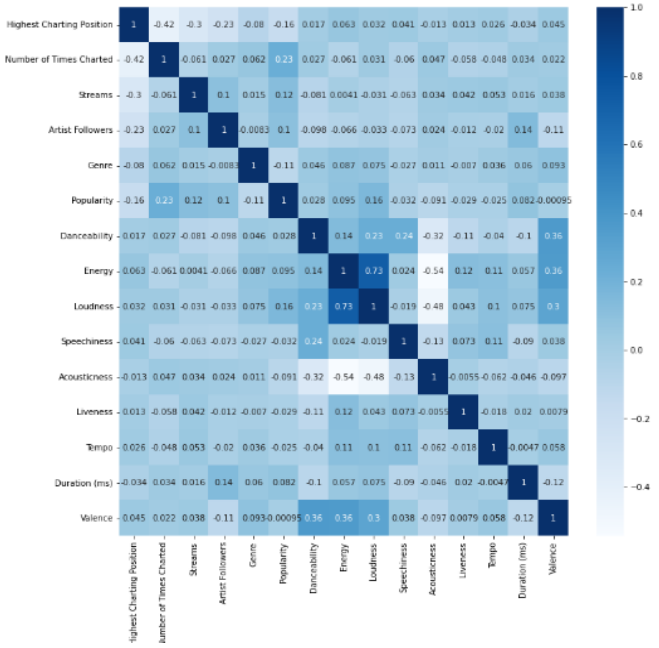

Fig. 6. Heat map of Correlation Plot

Figure 7 shows that Dance Monkey is the most charted song while Beggin' is the least charted song.

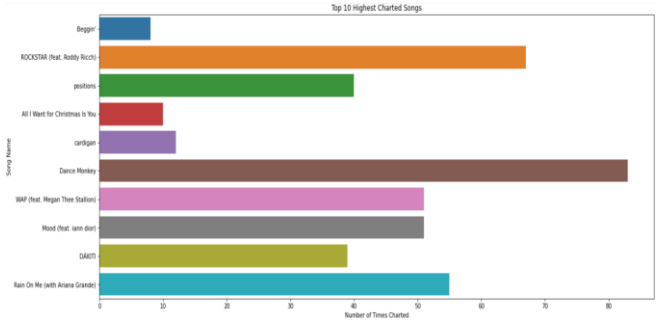

Fig. 7. Histogram of number of times charted Vs Song Name

Figure 8 shows that the artists, Maneskin have the highest streams for their song while Bad Bunny has the lowest.

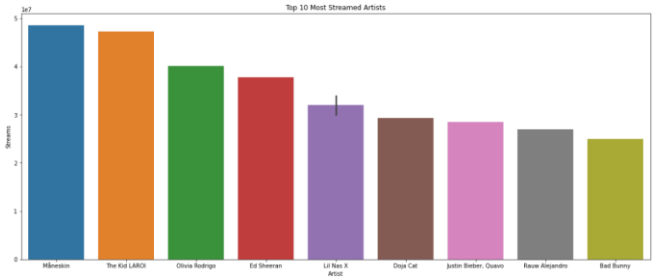

Fig. 8. Histogram of Artist Vs Streams

Figure 9 shows that the genres, Indie rock Italiano have the highest streams while trap latino has the lowest.

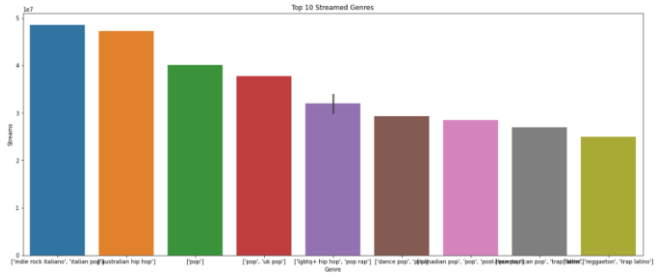

Fig. 9. Histogram of Genre Vs Streams

4. CONCLUSION

Spotify is a very large platform that is used around the world every day by millions of users and thus this research project helped us to gain more knowledge about it. It was insightful to know about the attributes that actually affect the streams of a song and what causes a user to listen to a particular artist. Moreover, we gained knowledge about various regression models which can be used for other projects in the future.

In this research paper, we have successfully predicted the popularity of songs using relevant attributes, which helped us gaining our desired accuracy. The most influential factors that fulfilled our objective to determine streams were loudness and energy. In the initial part of our research, we have used techniques like Lasso Regression, Ridge Regression, Random Forest and Linear Regression. We have obtained the best results from Random Forest, which came out to be 97.48%, Hence the most accurate regression model was Random Forest.

In the future, the research work can be improved with the help of a dataset with more songs and attributes to gain better accuracies of the models.

REFERENCES

[1] Dataset:https://www.kaggle.com/datasets/sashankpillai/spotify-top-200-charts-20202021

[2] J. S. Gulmatico, J. A. B. Susa, M. A. F. Malbog, A. Acoba, M. D. Nipas and J. N. Mindoro, "SpotiPred: A Machine Learning Approach Prediction of Spotify Music Popularity by Audio Features," 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 2022, pp. 1-5, doi: 10.1109/ICPC2T53885.2022.9776765. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[3] Araujo, C. S., Cristo, M., & Giusti, R. (2019). Predicting music popularity on streaming platforms. Anais Do Simpósio Brasileiro De Computação Musical (SBCM 2019).

Available: https://doi.org/10.5753/sbcm.2019.10436

[4] Suta, M. (2018, January 1). Hitpredict: Predicting hit songs using Spotify Data Stanford Computer Science 229: Machine learning. Academia.edu. Retrieved October 7, 2022.

Available: https://www.academia.edu/73249006/Hitpredict_Predicting_Hit_Songs_Using_Spotify_Data_Stanford_Computer_Science_229_Machine_Learning

[5] Nijkamp, R. (n.d.). Prediction of product success: Explaining song popularity by audio features from Spotify data. Retrieved October 6, 2022.

Available: https://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf

[6] https://www.statisticshowto.com/lasso-regression/

[7] https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

[8] https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[9] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[10] https://www.mathworks.com/help/stats/what-is-linear-regression.html