iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38702

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Big Data Analytics: A Spotify Case Study

Suraj Ingle[1], Jesica Shah[2], Ruchi Mehta[3]
[1]*Indian Maritime University (Visakhapatnam)*
[2]*Usha Pravin Gandhi college of Arts, Science and Commerece*
[3]*Veermata Jijabai Technological Institute*

Abstract: ***By developing products that are in line with consumer needs, anticipating their profitability and manufacturing them, Big Data has opened up a lot of possibilities for building customer loyalty and commercial business by proactively engaging and comprehensively streamlining offers across all customer touch points. The use of big data to determine the best, most efficient ways to engage and interact with their customers will be discussed in this paper. An insight into how Spotify intends to provide music lovers additional ways to find their favourite songs, interact with artists, and improve Spotify recommendations has been provided.***

Keywords: Big Data, Data Analytics, Customer Satisfaction, *Exploratory Data Analysis*

I. INTRODUCTION

In global firms' new product development (NPD) endeavours, big data is becoming increasingly important. Firms are increasingly relying on valuable knowledge obtained from big data to be competitive in today's fast-changing business climate (Barton 2 and Court, 2012) [1] ;(Salehan and Kim, 2015) [2]. As a result, companies are increasingly turning to big data to a) better understand their customers, b) produce better products, and c) provide more personalised services to their customers. One of the most significant potential benefits of collecting big data, according to Davenport (2012) [3], is its application in the development of new products and services. However, just a few studies have looked into how firms may improve their service using big data. This article explains how firms may use big data to reduce time to market, increase user adoption, and cut costs while developing new products. Big data enables companies to have a better understanding of their products, customers, and markets, which is essential for consumer loyalty. Businesses' main challenge is figuring out how to use big data to improve customer intelligence. Whereas almost everyone, including marketers, corporate managers, researchers, and policymakers, have experienced problems and challenges as a result of the "big data" phenomenon: How can big data be used to benefit marketing, management, and policy-making? While 63 percent of organisations see big data analytics as a competitive advantage, 80 percent of marketers say they don't know how to turn data into action, and 95 percent of data within organisations remains unused, according to several academic and industry reports (Kiron et al., 2011) [4]; Rogers and Sexton, 2012) [5]; (Monetate, 2014 a,b) [6]. Even more baffling, according to one survey (Allen et al., 2005) [7], while 80% of CEOs believe they provide exceptional customer service, just 8% of customers concur. To demonstrate how the principles might be utilized to get additional benefit from big data, a case study with Spotify, a commercial music streaming service firm, is used. The repercussions for practitioners and academia are examined, and conclusions are offered towards the end.

II. A CASE STUDY ON SPOTIFY

Spotify is a music streaming service that is accessible online. Most of the data is centred on the user, allowing them to make music recommendations and select the next songs. They do everything they can to incorporate the culture into every decision and activity. Spotify aims to be completely data-driven in its data analytics. Spotify employs the complete collaborative filtering process, providing it with the most up-to-date representation of users and artists. Spotify data analytics primarily deals with massive amounts of data, which is nothing new, and also take into account the link between data warehouses and data marts. The data analysis gives us a better understanding of listening patterns and preferences.

A. Why Spotify makes use of Big Data-

Spotify takes a critical approach to utilising the data collected by its users by using it to create material that each user will regard as unique to their preferences. The goal is to provide users with a positive experience that will lead to them becoming loyal clients. Various Artificial Intelligence and Machine Learning algorithms have been used to do this.

B. Developing Personalized Content

The platform's "Discover" feature, for example, which was first introduced in 2012, plays a critical part in Spotify's data collection. This feature started out as a playlist of music published by the user's favourite artists, but it evolved into a recommendation engine that suggested a collection of tracks as the user's playlist progressed, all of which were aligned along the lines of the songs in the playlist. In order for the platform to be able to personalise these playlists, it has to pay close attention to both the tracks that users stream and how they engage with each tune in general.

C. For Enhanced Marketing through Targeted Ads

While enhancing the customer experience, Spotify has also been willing to incorporate a massive amount of data provided by its users for the sake of upgrading their ad campaigns and better targeting their customers. This is done by the platform reviewing the information they've gathered about their listeners and then using that information to develop commercials that are specifically targeted at the platform's target audience.

D. Continuously Updating its System

The freedom to explore the site's well-known playlists helps the platform to generate data from an additional hundred million or more users, which is particularly beneficial as the firm focuses on improving its suggestion algorithms to provide a satisfying tailored experience to its customers. In order to make their vast quantity of data available to their musicians and managers, Spotify launched a Spotify for Artists tool, which gives them access to information like which playlists have been helping them attract new users and the total number of streams they've received

III. ANALYSIS

For our analysis, we will analyse a large Spotify dataset in order to find patterns and decipher whether there is a fixed recipe or a common criterion to produce hit songs. The dataset used consists of over 17k rows with multiple parameters as columns.We will conduct an Exploratory Data Analysis (EDA) on a specified Spotify music dataset in order to derive our conclusions.

The initial dataset has the following key attributes:

Range Index: 174389 entries,

Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	acousticness	174389 non-null	float64
1	artists	174389 non-null	object
2	danceability	174389 non-null	float64
3	duration ms	174389 non-null	int64
4	energy	174389 non-null	float64
5	explicit	174389 non-null	int64
6	id	174389 non-null	object
7	instrumentalness	174389 non-null	float64
8	key	174389 non-null	int64
9	liveness	174389 non-null	float64
10	loudness	174389 non-null	float64
11	mode	174389 non-null	int64
12	name	174389 non-null	object
13	popularity	174389 non-null	int64
14	release date	174389 non-null	datetime64[ns]
15	speechiness	174389 non-null	float64
16	tempo	174389 non-null	float64
17	valence	174389 non-null	float64
18	year	174389 non-null	int64
19	release_year	174389 non-null	int64
20	release_month	174389 non-null	object

However, since the **id** column won't be that useful for our analysis, we can drop it.
Here is a brief glance at the refined dataset being analysed –

	name	popularity	duration_ms	explicit	artists	release_date	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	release_year	release_month
430239	Árla (Cantilenas De Bachianas Brasileiras No. 5)	18	269600	0	['Quinteto Armorial']	1978-08-05	0.490	0.249	5	-13.001	1	0.0487	0.951	0.743000	0.1080	0.277	150.677	4	1978	August
139556	Trouble and Me	17	134547	0	['Stonewall Jackson']	1964-01-01	0.710	0.585	3	-9.674	1	0.0467	0.530	0.000000	0.1240	0.961	98.049	4	1964	January
242923	For What It Was - Radio Edit	38	327030	0	['Shriock']	1996-04-04	0.723	0.775	11	-8.345	0	0.4350	0.592	0.000000	0.0672	0.742	89.080	4	1996	April
355207	Mogu li reci, ne volim te	18	190973	0	['Dragana Mirkovic']	1988-01-01	0.501	0.515	4	-9.454	0	0.0736	0.315	0.000033	0.0485	0.239	187.944	3	1988	January
161846	Balladen om Fredrik Åkare	13	142653	0	['Cornelis Vreeswijk']	1964-01-01	0.566	0.143	6	-17.104	0	0.1710	0.953	0.000000	0.1040	0.635	85.633	3	1964	January

a) We first perform some basic data cleaning methods such as removing any duplicate entries or empty rows before we start our main analysis. Now, as part of our main data exploration and analysis, we look at the distribution of these tracks by popularity in Fig. 1:

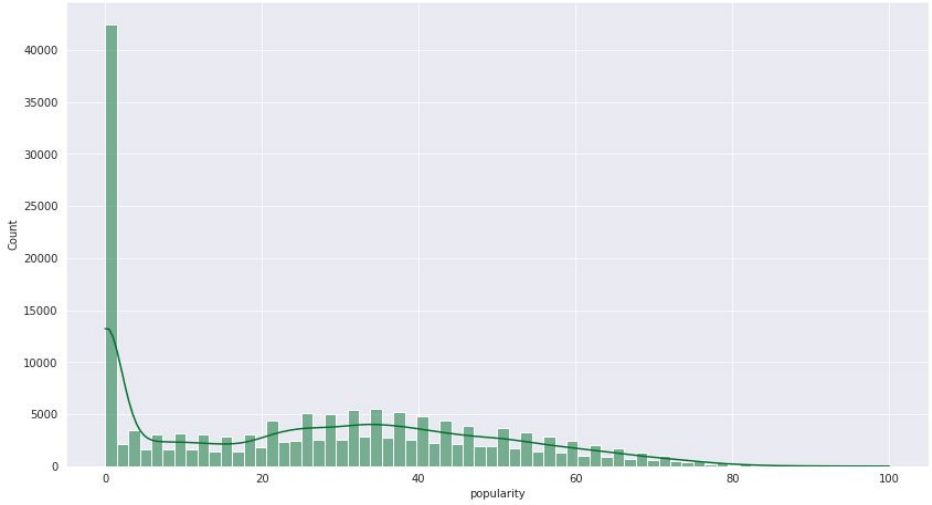


Fig. 1 Distribution of tracks by popularity

b) From the graph in Fig.2. it becomes evident that more than 45k songs are in a popularity graveyard. And most of the songs are distributed between 1 and 40 points of popularity approximately. This indicates that the music market is highly competitive in this time range. We now try to analyse past trends as well to see if it was always competitive and whether this trend has been increasing over the years.

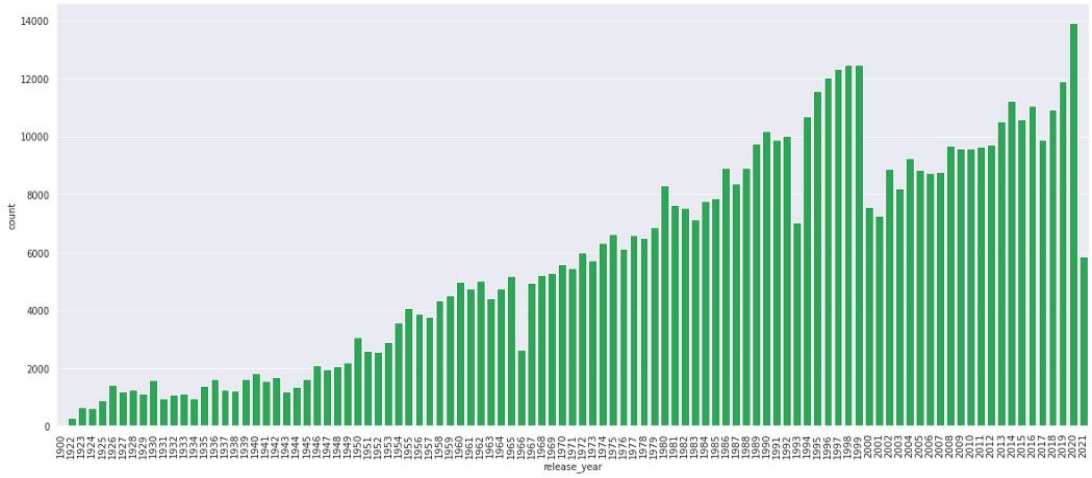


Fig. 2 Distribution of tracks produced over time

Based on this graph in Fig.2. it seems it's getting more and more competitive year after year, with nearly 140k songs produced in 2020. This could be attributed to the fact that 2020 was a year where many people had a lot of free time at home and hence this led to a boom in the music generation segment.

If we explore the overall metrics in our dataset, we get the following observation:

	popularity	duration_ms	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	release_year
count	585203.000000	5.852030e+05	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000	585203.000000
mean	27.628155	2.300585e+05	0.044106	0.563641	0.542274	5.221689	-10.201890	0.658840	0.104924	0.449564	0.112913	0.213979	0.552456	118.486428	3.873509	1988.597639
std	18.346391	1.264305e+05	0.205331	0.165980	0.251744	3.519364	5.078418	0.474099	0.180070	0.346596	0.266257	0.184365	0.257625	29.754339	0.472655	22.771278
min	0.000000	3.344000e+03	0.000000	0.000000	0.000000	0.000000	-60.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1900.000000
1%	0.000000	5.733300e+04	0.000000	0.159000	0.033600	0.000000	-26.159000	0.000000	0.025100	0.000058	0.000000	0.038200	0.039700	63.849040	1.000000	1928.000000
10%	2.000000	1.342800e+05	0.000000	0.335000	0.190000	0.000000	-16.878000	0.000000	0.029400	0.010500	0.000000	0.072500	0.189000	81.302000	3.000000	1956.000000
20%	9.000000	1.649512e+05	0.000000	0.421000	0.298000	2.000000	-13.879000	0.000000	0.032400	0.057600	0.000000	0.090800	0.299000	91.801000	4.000000	1969.000000
30%	16.000000	1.836400e+05	0.000000	0.482000	0.390000	2.000000	-12.035000	0.000000	0.035600	0.146000	0.000000	0.105000	0.390000	99.812000	4.000000	1979.000000
40%	22.000000	1.991868e+05	0.000000	0.532000	0.472000	4.000000	-10.549000	1.000000	0.039300	0.271000	0.000003	0.117000	0.482000	108.505800	4.000000	1986.000000
50%	27.000000	2.149600e+05	0.000000	0.577000	0.549000	5.000000	-9.242000	1.000000	0.044300	0.422000	0.000024	0.139000	0.564000	117.424000	4.000000	1992.000000
60%	33.000000	2.322270e+05	0.000000	0.620000	0.630000	7.000000	-8.074000	1.000000	0.051600	0.579000	0.000215	0.175000	0.647000	125.014000	4.000000	1998.000000
70%	38.000000	2.521600e+05	0.000000	0.663000	0.708000	7.000000	-7.003000	1.000000	0.064800	0.721000	0.002390	0.237000	0.727000	132.016400	4.000000	2004.000000
80%	44.000000	2.779600e+05	0.000000	0.709000	0.793000	9.000000	-5.952000	1.000000	0.095200	0.842000	0.042500	0.318000	0.811000	141.071000	4.000000	2010.000000
90%	52.000000	3.234930e+05	0.000000	0.769000	0.881000	10.000000	-4.755000	1.000000	0.213000	0.949000	0.644000	0.428000	0.899000	160.182800	4.000000	2016.000000
99%	71.000000	5.976964e+05	1.000000	0.893000	0.979000	11.000000	-2.548000	1.000000	0.950000	0.995000	0.943000	0.936000	0.969000	195.367940	5.000000	2020.000000
max	100.000000	5.621218e+06	1.000000	0.991000	1.000000	11.000000	5.376000	1.000000	0.971000	0.996000	1.000000	1.000000	1.000000	246.381000	5.000000	2021.000000

Fig. 3 Analysis of the dataset

From this table in Fig.3., it's evident that there's a big leap between the 90% and 99% percentiles in the popularity variable, compared to the previous ones. So, it seems that a few great hits are close to scoring 100 in popularity. This means that there is a select group of tracks being quite popular on Spotify

We can therefore analyse if it is possible to get there by putting the right chords and rhythm into our song? We plot a correlation chart (heatmap) to find this out:

The attributes taken into consideration are -

track attributes = ["popularity","acousticness","danceability","energy","duration_ms","instrumentalness", "valence", "tempo", "liveness", "loudness", "speechiness"]

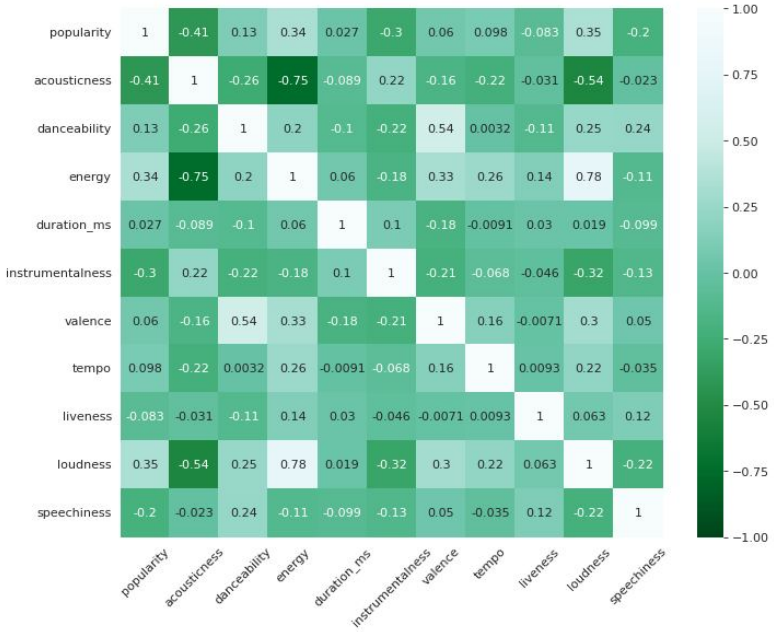


Fig. 4 Correlation chart (heatmap)

Based on this heatmap in Fig.4., we can see that there are no significant correlations between popularity and the track's attributes. Still, it would be worth diving deep into the three attributes that showed a positive correlation: danceability, energy and loudness in Fig.5.

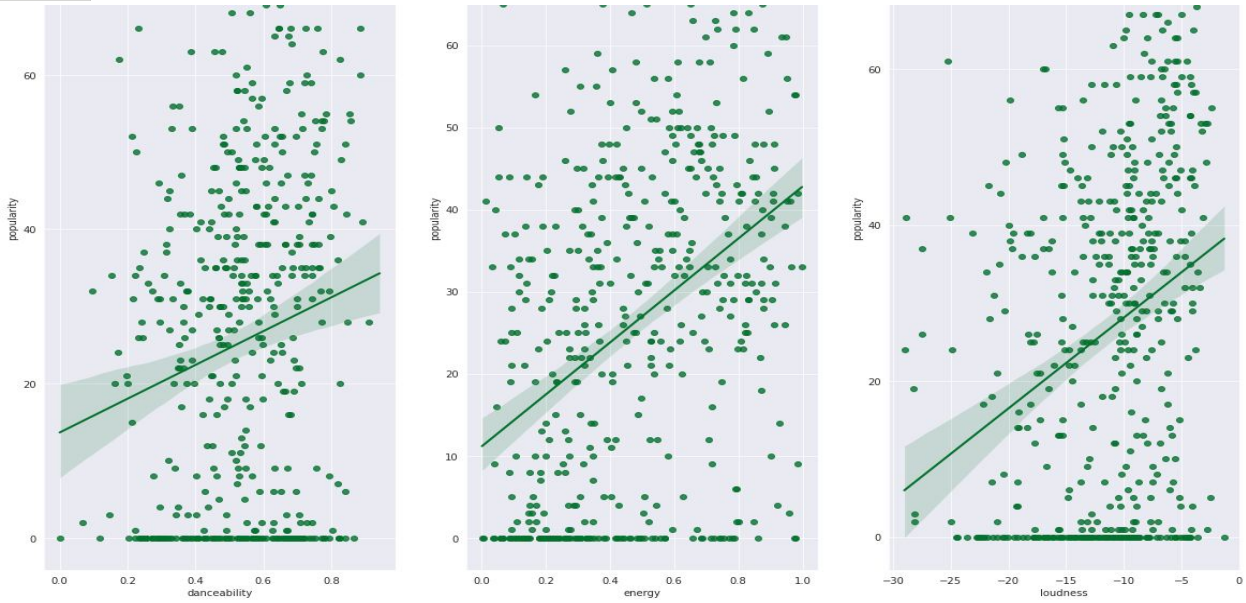


Fig. 5 Correlation chart of danceability, energy and loudness

This is confirming what we first saw in our correlation heatmap in Fig.4, but revealed something quite interesting for our analysis: most of the high popularity outliers are found within the highest ranges of the three attributes, especially for loudness. This might be a huge stepping stone towards solving our main question.
It would be best to subset our data to get the most popular songs, so we can see how present are these attributes:
So, we gather a subset of all songs that have a popularity factor of greater than or equal to 80

For this new subset, we can see in Fig.6, what they have in common by plotting their attributes by mean:

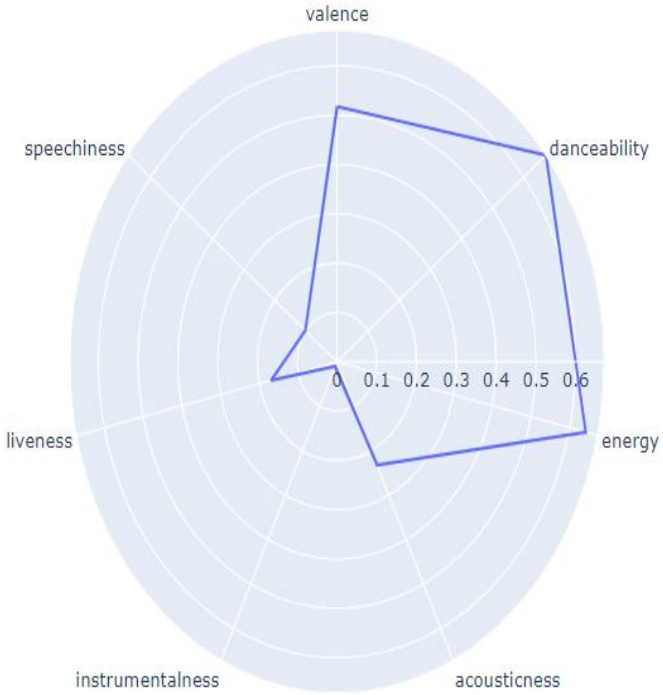


Fig. 6 Plot of different attributes by mean

From the graph it seems, danceability is a quite strong contestant as it is a major attribute in most popular songs as seen in Fig7.We can also take a look at how these have been part of the most popular songs over time:

Fig. 7 Audio attributes over time for popular songs

Seems like popular songs have always been quite energetic, danceable and loud during the last 50 years, as well as happy. This looks like a clear indicator of people's music taste. Let's take a look at the top 25 artists with songs that are currently popular and compare it with a random sample

['Taylor Swift'] 14
['Billie Eilish'] 10
['BTS'] 9
['Bad bunny'] 9
['Ariana Grande'] 9
['XXXTENTACION'] 8
['Ed Sheeran'] 6
['Lewis Capaldi'] 6
['Harry Styles'] 6
['Pop Smoke'] 5
['One Direction'] 5
['Dua Lipa'] 5
['Imagine Dragons'] 5
['Sam Smith'] 4
['Morgan Wallen'] 4
['Travis Scott'] 4
['Post Malone'] 4
['The Kid LAROI'] 4
['Bruno Mars'] 4
['Justin Bieber'] 4
['Arctic Monkeys'] 4
['Miley Cyrus'] 3
['Shawn Mendes'] 3
['Coldplay'] 3

Seems like Taylor Swift, BTS and Billie Eilish are quite popular right now with >9 current hit tracks. However, all of the artists in the top 25 list are already very well known, as well as many of the ones in the list.
So, it can be concluded as from the analysis of this data that:

Producing a hit track won't necessarily depend just on how happy, energic, danceable or loud your song is, but more likely it would be related to your current popularity as an artist. However, in order to increase the probability of generating a relatively popular song, the analysis suggests that it would be a good idea to add attributes that have high popularity like the ones shown in the analysis above.

IV. LIMITATIONS

A. The dataset used in the current analysis is not large enough to capture significant trends at the root level. It is suitable for a high-level exploratory analysis but not big enough to derive business insights into the data.
B. User data is inherently biased as it only captures the data of a specific section/group of people. Hence the analysis performed may be affected and the insights captured might possess a certain level of bias that corrupts the derivations.
C. Some parameters used to analyse data may not be significant factors in deriving the required conclusions and insights. Considering them might skew the results by some margin.
D. Most of the big data captured from users is prone to noise. A single outlier may have a negative effect on the whole dataset.
E. Analysing large quantities of data may slow down systems and also have a significant cost associated with it. Sometimes this cost outweighs the gains achieved by conducting a thorough analysis.
F. Big data analysis is prone to security vulnerabilities. It can act as a potential point of failure or attack in cases of sensitive data.
G. Big data analytics cannot be conducted conclusively for data which has privacy restrictions.

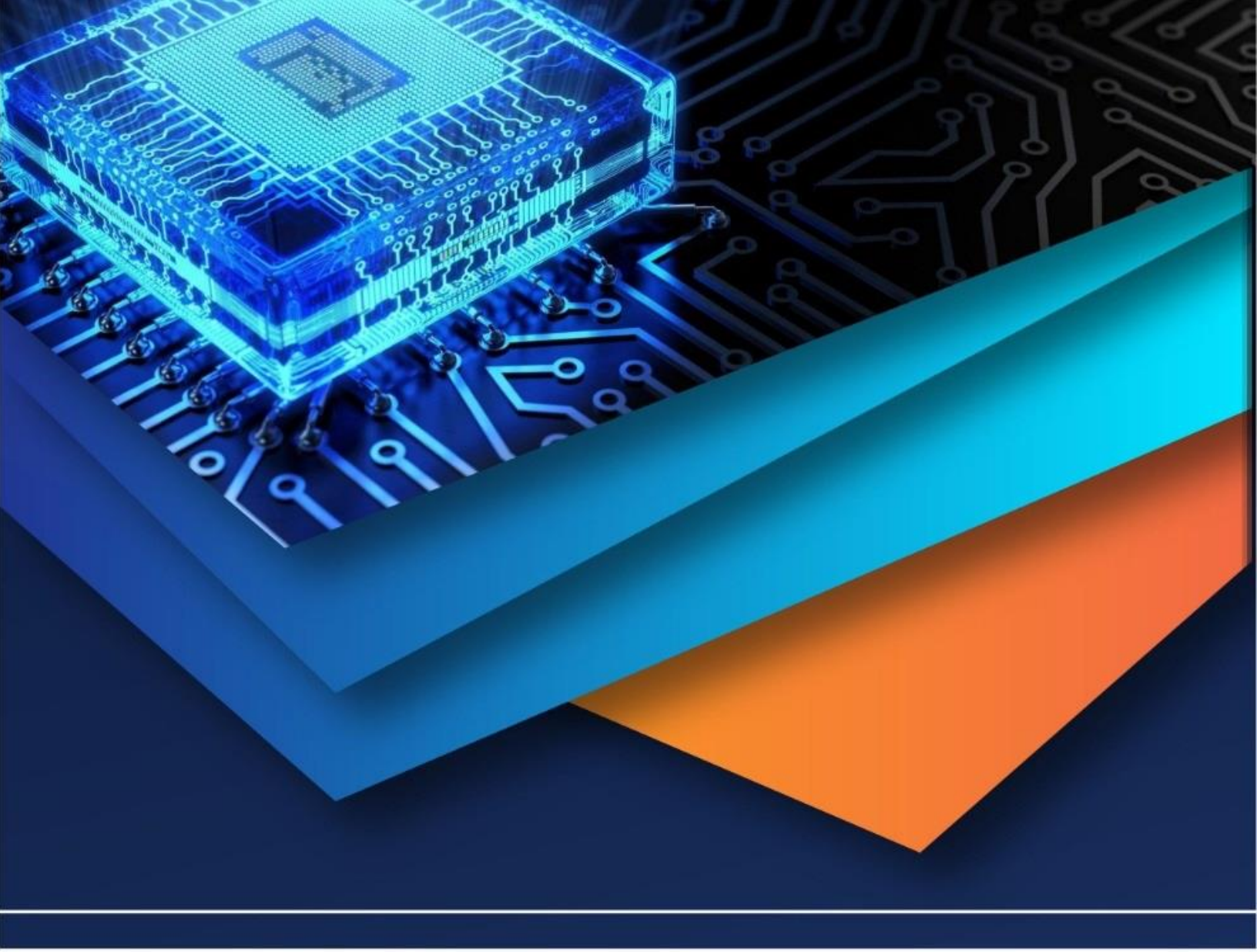
V. CONCLUSION

In today's environment, when streaming music has surpassed purchased music, the music industry has been forced to shift its focus away from record sales and toward collecting data with the purpose of determining the impact a particular song, artist, or album has on the general population. Because the data also provides a deeper understanding of listening trends, audience markets, and other areas, it is a never-ending revolution for those in the industry.

Spotify becomes an inadvertently self-marketable platform because users promote their engagement on their own accord because it is a social and sharing experience. By combining its application of data with a robust user experience custom made for social media, Spotify becomes an inadvertently self-marketable platform. Spotify would not have turned out the way it did if it hadn't been for big data. With a rising presence in numerous countries and a growing audience, more data will be generated in the future years. More data will result in better suggestions, better predictions, more users, and, as a result, more compensation to the rights holders. Spotify was able to completely transform the music industry because to big data.

REFERENCES

[1] Barton, D. and Court, D. (2012), "Making Advanced Analytics Work For You", Harvard Business Review, Vol. 90 No. 10, pp. 78-84.
[2] Salehan, M. and Kim, D. J. (2016). "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics", Decision Support Systems, Vol. 81, pp. 30- 40.
[3] DAVENPORT, T. H. 2014. Big data at work: dispelling the myths, uncovering the opportunities, Boston, Harvard Business Review Press.
[4] D. Kiron, R. Shockley, N. Kruschwitz, G. Finch, and M. Haydock. Analytics: The widening divide. MIT Sloan Management Review, 53(3):1–22, 2011.
[5] Rogers and D. Sexton. Marketing roi in the era of big data: The 2012 brite and nyama marketing in transition study. Technical report, Columbia Business School, http://www.iab.net/media/file/2012-BRITE-NYAMA-Marketing-ROI-Study.pdf, 2012.
[6] Monetate. Connecting data to action. Technical report, Monetate, 2014b.
[7] Allen, F. F. Reichheld, B. Hamilton, and R. Markey. Closing the delivery gap. Technical report, Bain and Company, 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY