BIG DATA ANALYSIS OF SPOTIFY MUSIC DATA

Team 17, GMU ECE552 Big Data Technologies, Fall 2022

Jaswanth Kumar Kunku
DAEN, GMU
skunku@gmu.edu

Abhishek Godavarthi
DAEN, GMU
agodavar@gmu.edu

Bhargav Teja Jakku
DAEN, GMU
bjakku@gmu.edu

Abstract: A growing number of automated systems enable users to collect, organize, and analyze massive amounts of web-based data through the distribution of digital music [1]. This study uses data obtained from different Spotify playlists through Kaggle repository and applies data cleaning techniques, exploratory and predictive analytics to uncover trends on how music interests are changing from 1921 to 2020. It also associates the relationships between audio characteristics and how it's affecting the song's popularity. The goal of the study is to identify and discover patterns in the data using relevant spark big data libraries available in python and accurately forecast the classification of highly popular songs from the less popular ones. In this study, we examine the data results in more detail, talk about the study's limitations, and present the key findings.

I. INTRODUCTION

Spotify is a proprietary Swedish audio streaming and media services provider founded on 23 April 2006 by Daniel Ek and Martin Lorentzon. It is one of the largest music streaming service providers, with over 456 million monthly active users, including 195 million paying subscribers, as of September 2022. Spotify offers digital copyright restricted recorded music and podcasts, including more than 82 million songs, from record labels and media companies. Users can search for music based on artist, album, or genre, and can create, edit, and share playlists [2].

"Historically, neither the creators nor the distributors of cultural products have used analytics - data, statistics, predictive modeling - to determine the likely success of their offerings. Instead, companies relied on the brilliance of tastemakers to predict and shape what people would buy" (Davenport et al.,2009). While tastemakers are still significant influencers of products also in cultural product markets, the way we consume is changing by technological developments [3].

People use music extensively in their daily lives, and as music becomes more digitized, huge quantities of musical data are created that tend to be further accumulated by music fans (Sloboda, 2011). This has caused music collections to expand beyond what was previously feasible, not only on the private shelf as audio or video CDs and domain discs, but also on the hard disk and Internet. With the development of new technology, it is now impossible for one person to keep up with the music and the connections between various tunes [4].

With the increase in the Online streaming services, music service providers are continuously looking for methodologies that enable seamless user experience through useful recommendations based on the user preferences and previous music charts. While each streaming service have own intelligence in delivering the right recommendations; Spotify's research interests include large-scale recommendation algorithms for music and more generally audio discovery, algorithms for audio search through voice and text, query analysis for effective query suggestion, query completion and search assistance, multilingual information retrieval for voice search, and ranking algorithm for revenue and music ads [5].

II. OBJECTIVE

The goal of this research is to leverage big data tools and their capabilities to analyze large amounts of musical data (~1.04million) and identify the patterns in the dataset.

Our in-depth analysis of this dataset has led us to the following research questions:

1. *How did music and its popularity has changed over the past decade?*
2. *Which genres and artists were most popular?*
3. *Which music genres are influencing the public most?*
4. *What proportion of songs released are Explicit?*

III. LITERATURE REVIEW

The subject of music success prediction has been receiving more attention for years and we highlight three general tactics: The first gauge's public opinion at the moment and forecasts how popular a song or album will be using information from social networks. The second predicts the success of other songs using acoustic data from previously popular songs. The third method makes predictions about future inclusion of a song in a certain chart based on historical data from charts.

Research from Lee et al. (2015) shows that it is feasible to predict the popularity metrics of a song significantly better than random chance based on its audio signal. Additionally, Ni et al. (2015) also showed that certain audio features such as loudness, duration and harmonic simplicity correlate with the evolution of musical trends [3].

Utilizing social networks and social data is another strategy that businesses utilize in research for recommendation and prediction. Research on the use of social media data, such as that from Twitter, is one of the methodologies used in HSS that look for social popularity measurements. When recent music charts are available, Zangerla et al. (2016) discovered that leveraging Twitter posts can be helpful in predicting future charts[6]. Similar to this, Kim's research demonstrates a strong association between Twitter user data on music listening behavior and music's chart success [7].

Additionally, Bischoff et al. (2009) present a model for predicting music popularity based on social interaction data from Last.fm, with encouraging outcomes. This approach's downside is that it necessitates a lot of data [8].

Some popular approaches in product success prediction like the use of prediction markets (Matzler et al., 2013) are not as relevant for new content as it is for products.

IV. DATA REVIEW

The dataset used for this study was retrieved from Spotify API and made the publicly available in Kaggle Repository in CSV format[9]. It consists of two files containing information about the artists and tracks from the year 1921 to 2020 with 1.16 million and 0.58 million records respectively.

There are 25 attributes [13] in the dataset. The attributes are:

- **id**: The unique identifier for the track.
- **Track name:** Name of the song
- **Track popularity:** A measure between 0 -100 where 0 indicates low popularity and 100 indicates high popularity.
- **duration_ms**: The duration of the track in milliseconds.
- **Explicit:** Explicit describes whether a song is suitable for age below 18 or not. 0 indicates suitable,1 indicates not suitable.
- **Artist Id:** The unique identifier for the artist.
- **Artist Name:** Name of the person who played the song
- **Release date:** The year and date on which the song was released.
- **Danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
- **Energy**: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- **Key**: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- **Loudness**: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.

Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.

- **Mode**: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- **Speechiness**: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **Acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
- **Instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal".
- **Liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- **Tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **Valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- **time_signature**: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- **Genre**: Type of the genre the song belongs to. For example, Pop, Latin etc.
- **Artist Popularity**: Measusre between 0 to 100
- **Followers**: Number of followers

Attribute	Pre-Defined	Converted
ID	string	string
Followers	string	double
Genre	string	string
artist_name	string	string
artist_popularity	string	float
track_name	string	string
track_popularity	string	double
duration_ms	string	int
explicit	string	int
artists	string	string
id_artists	string	string
release_date	string	date
danceability	string	double
energy	string	double
key	string	int
loudness	string	double
mode	string	int
speechiness	string	double
accousticness	string	double
instrumentalness	string	double
liveness	string	double
valence	string	double
tempo	string	double
Time_signature	string	string

Table 1: Dataset Attributes

V. SYSTEM ARCHITECTURE AND METHODOLOGY

For this analysis we used windows Azure VM with Spark cluster running on it along to manage the enormous dataset. All the analysis and modelling are done using jupyter notebook with relevant pyspark libraries and data ingestion using MongoDB respectively. The following figure depicts the overall system architecture.

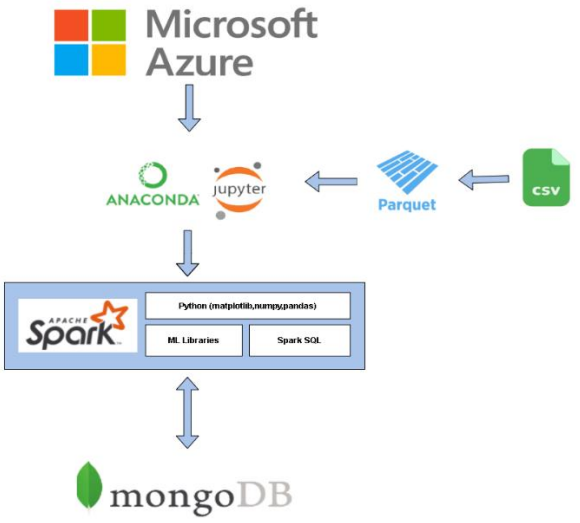

Fig 1: System Architecture

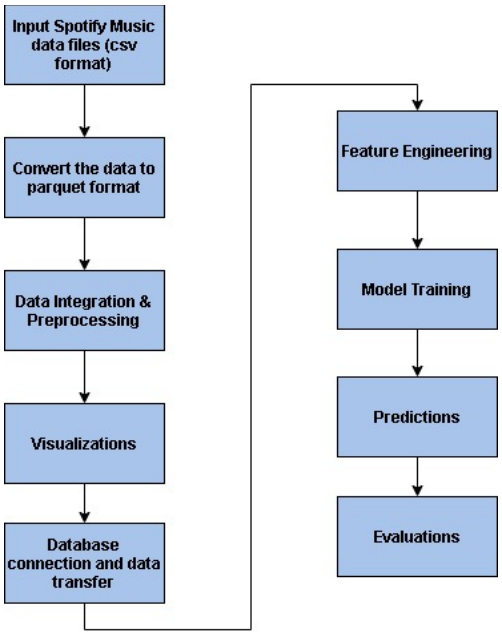

Fig 2: Data Flow Diagram

As illustrated in Figure 2, the project's structure can be split into nine steps. Initially, the datasets consisting of artists and song information are loaded into jupyter notebook. Next, the data obtained from the two csv files are converted into parquet format for fast processing. Upon conversion data processing methodologies like data merging, missing values and duplicates handling tasks were performed. Next, the cleaned data is ingested into MongoDB using spark connector as instructed to demonstrate the knowledge of data storage. Once preprocessed, exploratory data analysis was performed and represented the outputs through visualizations. Finally feature engineering methods such as String Indexing, Vector Assembler are used to transformed data for model training and evaluation.

In this study, we presented a few predictive analytics methods for developing ML algorithms. The following are :

- **Decision Trees**

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter[10].

- **Random Forest**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. [11].

- **Logistic Regression**

It is used for predicting the categorical dependent variable using a given set of independent variables[12].

Performance Evaluation Model

The evaluation of Machine learning algorithm's performance involves testing the proposed models. Different Performance Evaluation parameters such as Accuracy and ROC Area are used to evaluate the performance of the ML models.

VII. DATA PREPROCESSING

As part of data preprocessing, the 2 csv files have been converted to parquet format and joined into a single dataframe with track_id using right join. As a next step, we have dropped few insignificant columns such as "id"," id_artists" and "time_signature". The data in "release_date" is in irregular format, so we

considered only release year of the song. Type casting of columns has been done as shown in table 1.

Next, we have identified the N/A values and replaced numerical attributes with the median of the corresponding variables found a column that has complete null values. We have also dropped categorical variables.

```
+---------+------+----------+---------------+---------+--------------+----------+--------+
------+-----------+-----------+------+---+--------+----+----------+----------+------------
---+--------+-------+-----+
|followers|genres|artist_name|artist_popularity|track_name|track_popularity|duration_ms|explicit|
artists|release_date|danceability|energy|key|loudness|mode|speechiness|acousticness|instrumentaln
ess|liveness|valence|tempo|
+---------+------+----------+---------------+---------+--------------+----------+--------+
------+-----------+-----------+------+---+--------+----+----------+----------+------------
---+--------+-------+-----+
|    116735|154519|    116735|         461627|       71|          1854|       826|     416|
71|       2965|      2286|   859|430|     268| 160|        91|        64|          45|
34|     27|   23|
+---------+------+----------+---------------+---------+--------------+----------+--------+
------+-----------+-----------+------+---+--------+----+----------+----------+------------
---+--------+-------+-----+
```

Fig 3: N/A Values

Finally, the data has been checked for duplicates and 950 duplicates have been dropped. After all the preprocessing the data with (431203 rows, 21 columns) is obtained and it is stored as a collection in MongoDB.

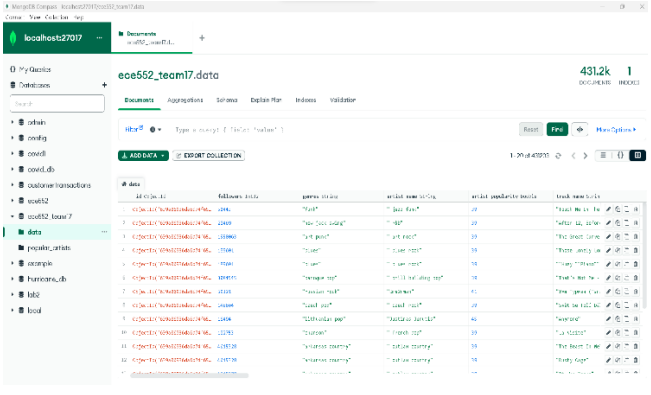

Fig 4: Mongo DB Data ingestion

VIII. Initial Findings and Results

Once the data has been completely preprocessed, we performed exploratory analysis on the data and identified key visualizations required for our analysis.

Identifying the most popular genre over years

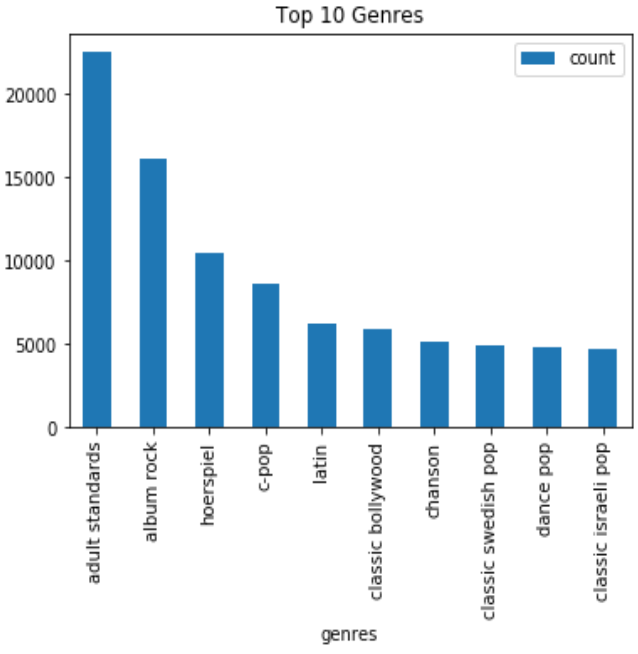

Fig 5: Popular Genres

The most popular genre based on Fig 5 is adult standards with 22529 songs followed by rock and hoerspiel with 16,188 and 10,423 songs respectively.

```
+---------------------+-----+
|genres               |count|
+---------------------+-----+
|adult standards      |22529|
|album rock           |16188|
|hoerspiel            |10423|
|c-pop                |8640 |
|latin                |6254 |
|classic bollywood    |5863 |
|chanson              |5170 |
|classic swedish pop  |4874 |
|dance pop            |4868 |
|classic israeli pop  |4752 |
|bossa nova           |4588 |
|alternative metal    |4190 |
|alternative rock     |3903 |
|bebop                |3900 |
|argentine rock       |3858 |
|j-pop                |3702 |
|classic italian pop  |3176 |
|classic icelandic pop|3111 |
|classic finnish pop  |3110 |
|art rock             |3065 |
+---------------------+-----+
only showing top 20 rows
```

Table 2: Popular Genre

Table 2 shows the count of popular genres over years.

Song popularity Trend over years

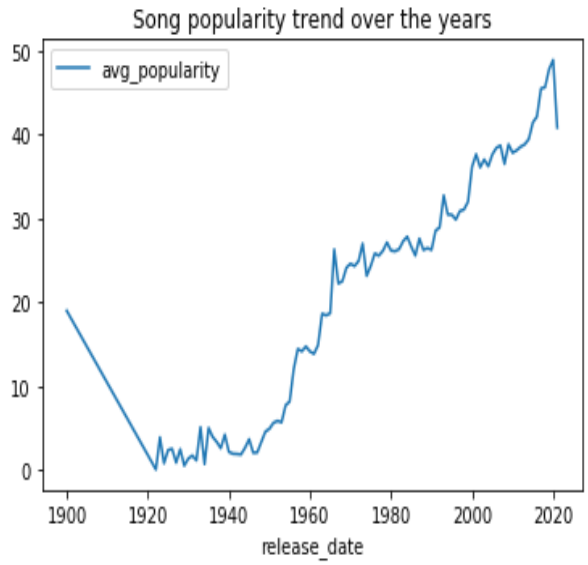

Fig 6: Song Popularity

Overall, the song popularity has always been on a rise only with an exception from the time period of 1900 to 1920(due to world war 1). From 1920 the popularity has increased exponentially and reached the peak in 2020.

release_date	avg_song_popularity
1900	19.0
1922	0.06
1923	3.9
1924	0.8
1925	2.38
1926	2.52
1927	0.88
1928	2.48
1929	0.52
1930	1.37
1931	1.72
1932	1.15
1933	5.11
1934	0.71
1935	5.02
1936	3.95
1937	3.35
1938	2.6
1939	4.22
1940	2.16

only showing top 20 rows

Table 3: Avg Song Popularity

Table 3 shows the average popularity of songs over years.

Trending song count over years

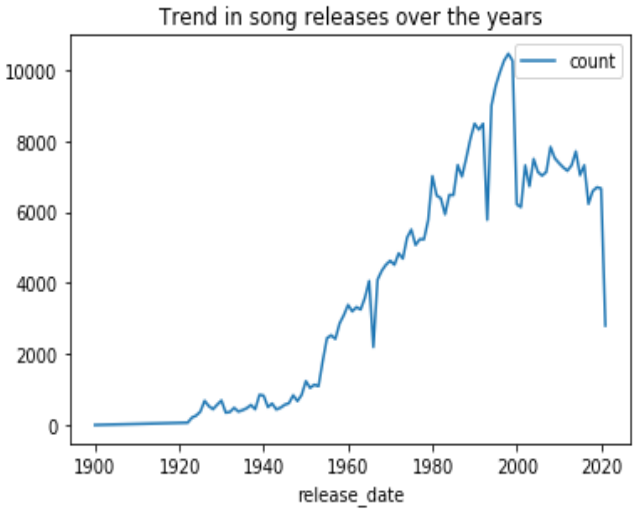

Fig 7: Trend in song releases over years

Based on Figure 5, we can see that the count of number of songs released in a year has also been increasing rapidly from 1900 to 2000. After 2000 there has been a slight decrease in the number of song releases when compared to previous years.

release_date	count
2021	2803
2020	6669
2019	6702
2018	6595
2017	6234
2016	7323
2015	7044
2014	7713
2013	7321
2012	7169
2011	7263
2010	7385
2009	7525
2008	7838
2007	7133
2006	7032
2005	7134
2004	7502
2003	6733
2002	7318

only showing top 20 rows

Table 4: Number of releases by year

Table 4 gives the number of song releases over years.

Most popular artists all time

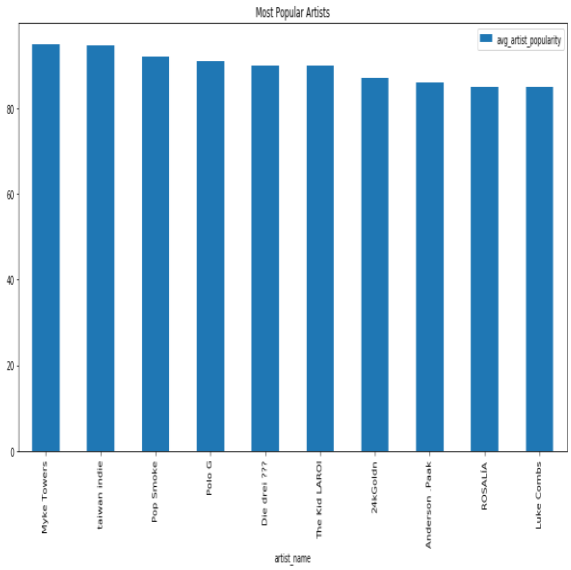

Fig 8: All time popular artists

The top 10 popular artists are Myke towers, Taiwan indie, Pop Smoke, Polo G, Die Drei, Kid Laroi, 24 Goldn, Anderson Paak and Luke combs and Rosalia and the average popularity for these artists is greater than 85.

artist_name	avg_artist_popularity
Myke Towers	95.0
taiwan indie	94.74
Pop Smoke	92.0
Polo G	91.0
Die drei ???	90.0
The Kid LAROI	90.0
24kGoldn	87.0
Anderson .Paak	86.0
Luke Combs	85.0
ROSALÍA	85.0

only showing top 10 rows

Table 5: Artists Average popularity

Genre with most followers

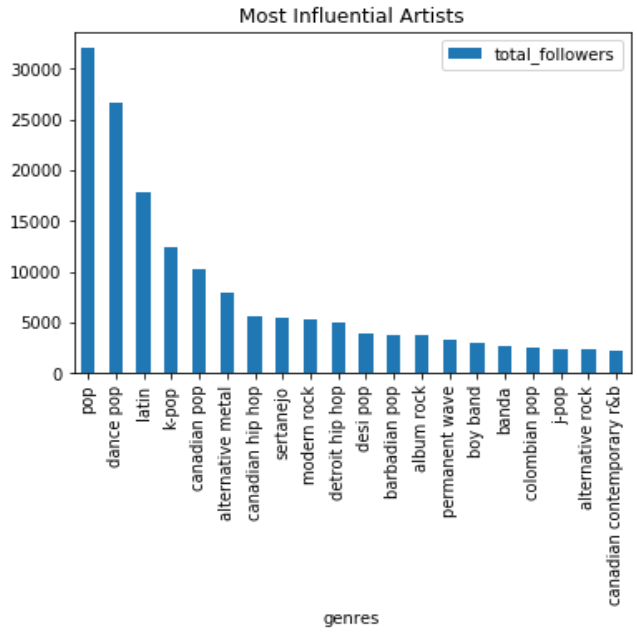

Fig 9: Popular genres by followers

The 10 most popular genres with most followers are pop, dance pop, latin, k-pop, Canadian pop, alternate metal, sertanejo, Canadian hip hop, desi pop and barbadian pop. The top 5 genres had more than 160,000 followers.

genres	total_followers
album rock	41976.098
pop	35563.598
dance pop	30367.63
classic rock	29630.115
latin	22734.799
adult standards	21366.607
k-pop	12386.771
alternative metal	12010.894
canadian pop	10950.572
beatlesque	10630.498
alternative rock	9624.755
classic bollywood	7723.6294
british invasion	6589.5234
hard rock	6329.127
sertanejo	6292.7007
detroit hip hop	6291.172
canadian hip hop	5651.537
modern rock	5486.1606
bolero	4805.8545
desi pop	4365.4526

only showing top 20 rows

Table 6: Popular genres by followers

Explicit Proportion of songs

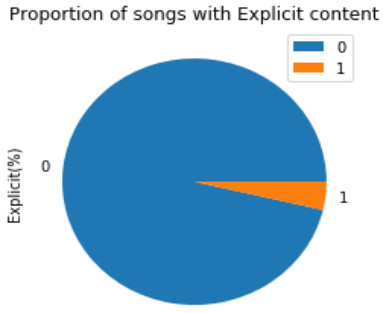

Fig 10: Explicit and Non-Explicit content proportion

From figure 10, it can be observed that most of the songs released over the years were not explicit songs. Only a few portions of the overall songs had explicit content in them. The explicit content constitutes of 3.67% and non-explicit constitutes of 96.33%

Machine Learning

Before training the classification models, the following feature engineering tasks were accomplished.

The data consists of one of the categorical column "genres" and it has been converted to numerical data by assigning labels to the input columns using String Indexer. Then all the numerical columns were captured, and a new data frame was created. The summary statistics of the data can be shown in the following figure.

summary	count	mean	stddev	min	max
followers	431203	1129717.646298843	3895894.0615830943	0	78900234
artist_popularity	431203	39.85238043334578	30.6338365222956	0.0	5566.0
duration_ms	431203	228047.42245299777	111314.82147273255	3344.0	4995083.0
explicit	431203	0.03677154379723703	0.18820063630360956	0.0	1.0
release_date	431203	1989.9592164247465	20.02446965643562	1900	2021
danceability	431203	0.5647943011100065	0.15876494682205539	0.0	0.9909999966621399
energy	431203	0.5629860435106407	0.2418301137583711	0.0	1.0
key	431203	5.234689925626677	3.51935406957673336	0.0	11.0
loudness	431203	-9.737432888568078	4.509405100504198	-60.0	5.375999927520752
mode	431203	0.6632792443466302	0.4725890459086408	0.0	1.0
speechiness	431203	0.09526686694673364	0.16629258756183968	0.0	0.9689999818801883
acousticness	431203	0.41759731390136207	0.33514968054230976	0.0	0.9959999918937683
instrumentalness	431203	0.08641783477533893	0.2318675068070202	0.0	1.0
liveness	431203	0.21400412114486947	0.18620873286747863	0.0	1.0
valence	431203	0.563026792514837	0.2526877619486645	0.0	1.0
tempo	431203	119.52783201510046	29.618647327941183	0.0	243.5070037841797
genresIndex	431203	235.5845344304191	402.7767154771507	0.0	3459.0
High_Popularity	431203	0.07271285218331042	0.25966487956116924	0	1

Table 7: Summary statistics of the data

Then all the numerical features have been plotted using a correlation matrix as shown in figure 11. It shows that feature "High popularity" which was obtained by transforming "track popularity" is highly correlated with followers, explicit, release date and loudness.

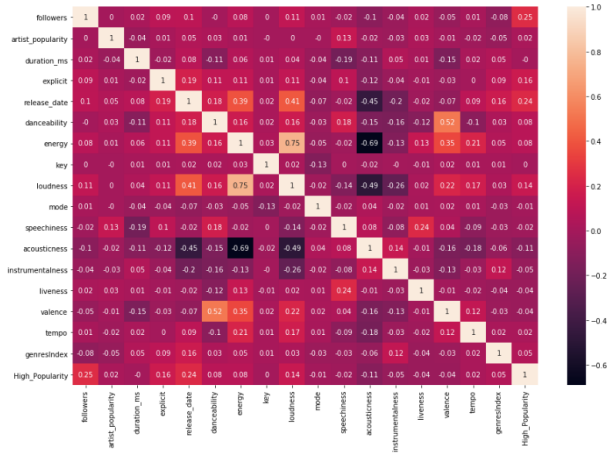

Fig 11: Correlation Matrix

The above variables seem to be significant in predicting popularity and have been converted into list of labels by fitting *Vector assembler* as input columns with the "High popularity" column as the output label.

For model building and evaluation, the data is portioned into two sets, by accommodating 70% of data for training and 30% of data for validation.

Logistic Regression

We used the logistic regression model [14] and then forecasted the results. Below is the ROC curve for logistic regression, which gave us an accuracy of 92.68.

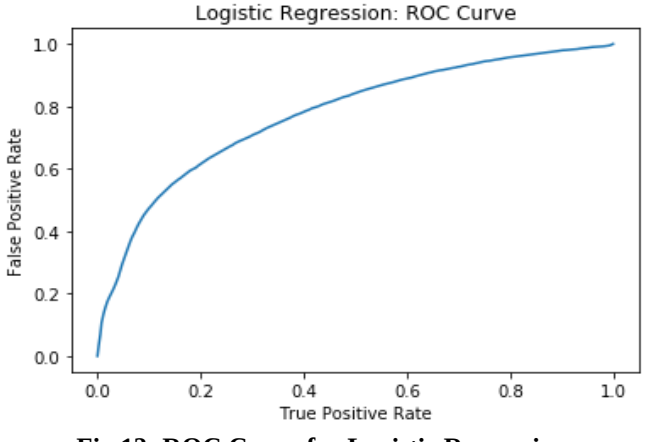

Fig 12: ROC Curve for Logistic Regression

Decision Trees

We used the decision tree model and then forecasted the results. The decision tree gave us an accuracy of 93.25 and ROC AUC of 0.19.

Random Forest

We used the random forest model and then forecasted the results. The decision tree gave us an accuracy of 93.05 and ROC AUC of 0.86.

Model	Accuracy (%)	AUC ROC
Logistic Regression	92.68	0.77
Decision Tree	93.25	0.19
Random Forest	93.05	0.86

Table 8: Model Summary

IX. CONLCUSION

Based on our analysis we were able to conclude that song popularity is increasing over the period, and it is a general phenomenon that the most recent songs are highly popular because of higher audience reach and thereby resulting in a declining trend for the previous years. When it comes to number of songs released over the year,21st century witnessed a sharp decline in the number of songs. When it comes to popular genres, pop and rock are widely supported by followers and it might grow on in the future. Finally, we also noticed there is very little proportion of explicit meaning in the songs which makes it ideal for listeners of various age categories.

To determine the best machine learning algorithm that is accurate and dependable, and finds the higher accuracy, we applied three machine algorithms to the Spotify Music dataset. The algorithms are Logistic Regression, Random Forest, and Decision Tree. We calculated, compared, and evaluated the various results obtained based on Accuracy and AUC ROC.

The Random Forest model outperformed all other algorithms after a precise comparison of our models revealed that it had higher efficiency and accuracy.

X. LIMITATIONS

The challenge of locating a completely balanced dataset is one of the key difficulties in this study. Since the data was compiled from a variety of user playlists and not everyone has access to all the music, most of the datasets connected to the study that are available online are unbalanced resulting in features with less correlation.

XI. ACKNOWLEDGEMENT

This research is the result of the support and guidance provided by Dr. Erton Bocci, George Mason University. We are extremely grateful to our professor for the professionalism, mentorship and expertise which proved critical throughout all stages of this project.

REFERENCES

[1] Sciandra, M., & Spera, I. C. (2020). A model-based approach to Spotify data analysis: a Beta GLMM. Journal of applied statistics, 49(1), 214–229. https://doi.org/10.1080/02664763.2020.1803810

[2] Wikipedia contributors. (2022, November 21). Spotify. Wikipedia. https://en.wikipedia.org/wiki/Spotify

[3] Nijkamp, R. (2018). Prediction of product success: explaining song popularity by audio features from Spotify data.

[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] Engineering, S. (2020, August 12). Search & Recommendations. Spotify Research. https://research.atspotify.com/search-recommendations/

[6] Zangerle, E., Pichl, M., Hupfauf, B., & Specht, G. (2016). Can microblogs predict music charts? An

analysis of the relationship between #nowplaying tweets and music charts.

[7] Kim, Y., Suh, B., & Lee, K. (2014). The future billboard: mining music listening behaviors of Twitter users for hit song prediction. Proceedings of International Workshop on social media

[8] Bischoff, K., Firan, C.S., Georgescu, M., Nejdl, W., & Paiu, R. (2009). Social knowledge-driven music hit prediction. Proceedings of International Conference on Advanced Data Mining and Applications, 43-54.

[9] Spotify Dataset 1921-2020, 600k+ Tracks. (2022, March 13). Kaggle. Retrieved December 11, 2022, https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks?select=tracks.csv

[10] Xoriant Corporation. (n.d.). Decision Trees for Classification: A Machine Learning Algorithm. Xoriant. https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm

[11] R, S. E. (2022, November 30). Understanding Random Forest. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[12] Logistic Regression in Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/logistic-regression-in-machine-learning

[13] Web API Reference | Spotify for Developers. (n.d.). https://developer.spotify.com/documentation/web-api/reference/

[14] GitHub - PacktPublishing/Machine-Learning-with-Apache-Spark-Quick-Start-Guide: Machine Learning with Apache Spark Quick Start Guide, published by Packy. (n.d.). GitHub. Retrieved December 14, 2022, from https://github.com/PacktPublishing/Machine-Learning-with-Apache-Spark-Quick-Start-Guide