Spotify popularity base on likes, nothing else matters□ ?

Popularity Index study, feature importances and prediction

DATA SCIENCE: PROJECT MACHINE LEARNING
July 2024
Author: María Fernández

What this presentation is about?

1. Spotify and music industry context
2. What data shows?
3. Machine learning applied
4. Predictions

Where words fail music speaks

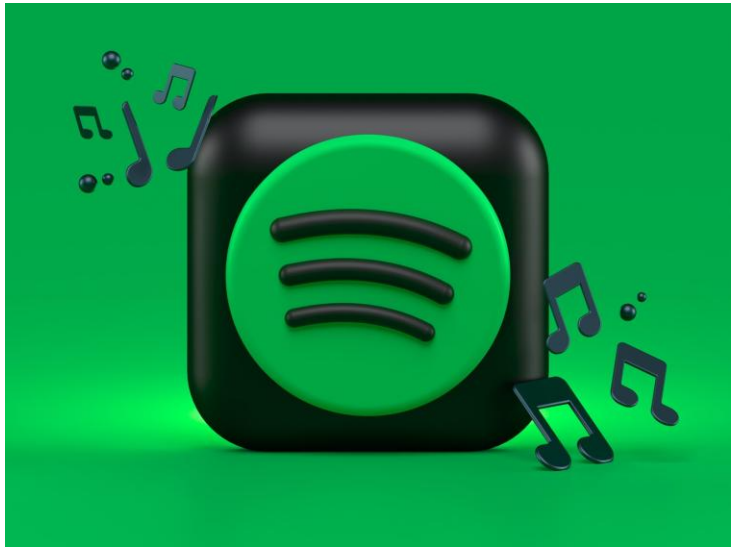
Spotify and music context

Spotify and music context

Songs's spotify popularity is related to several factors, artist and producers are interested in revealing formula to achieve success.

In Spotify words, Popularity Index, is a dynamic metric. This score, which ranges from 0 to 100, takes into account factors such as:

- **Recent plays**: How many times a song or album has been played in a **recent period** (usually 30 days).
- **Listener interaction**: including skips, repeats, and saves, affects the score.
- **Freshness of streams**: emphasizing the importance of current popularity.
- **Global vs. local popularity**: score is based on global plays, local popularity can also impact.



Keep in mind that it's a dynamic metric that can change over time based on user behavior and trends.

Therefore marketing promotion is a key point to achieve positions in Popularity Index, nothing else matters ?? (#Metallica)

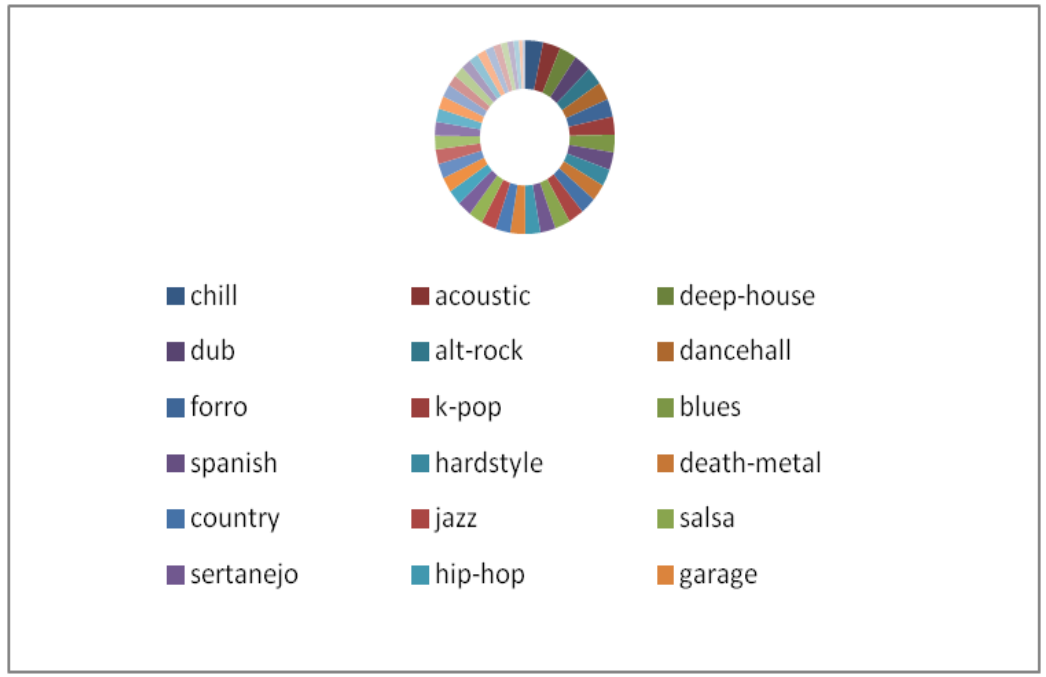
Where words fail music speaks

What this data set reveals?

What this data set reveals?

Pleanty of data

□ □ A sea of music and features to understand. and trends. □ □



What this data set reveals?

Features & Targets

□ □ A**TARGET: Popularity Spotify tracks' prediction**. We have a wide range of genres, years and artist. We want to analyse how different features influence the popularity.

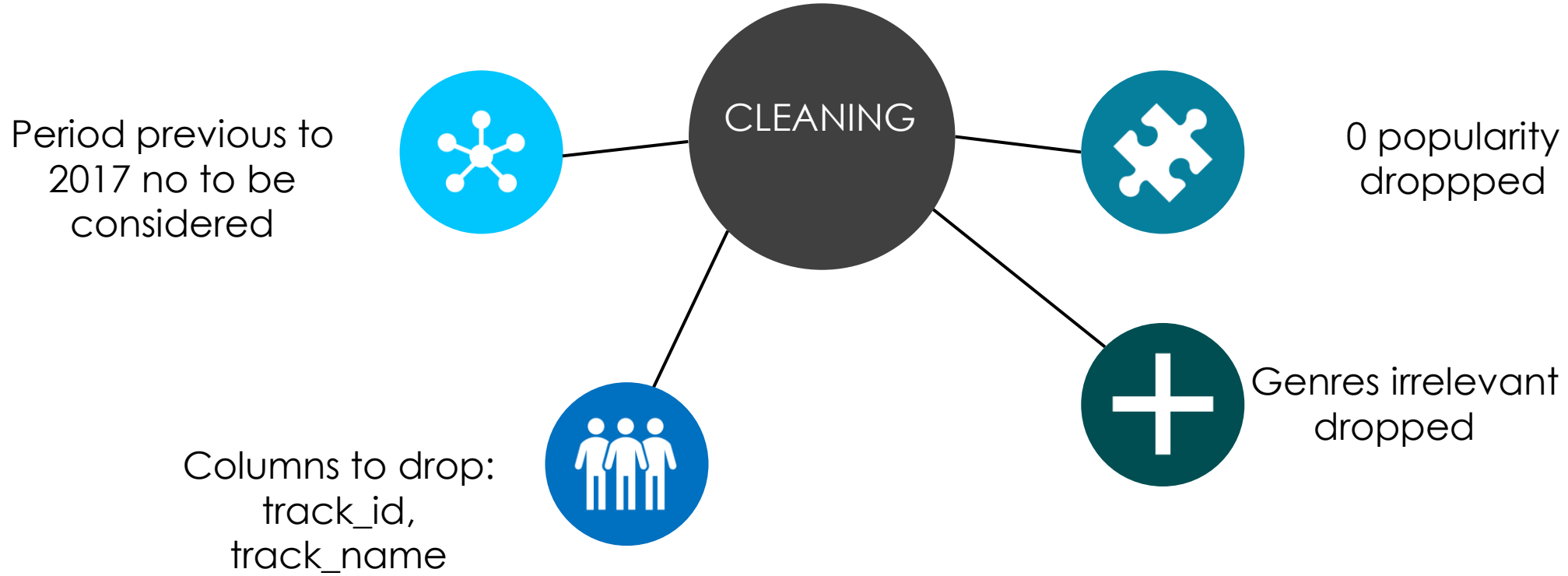
INITIAL HYPOTHESIS:

- Popularity is higher for certain artists.
- Genres with higher popularity are those related to more commercial music.
- Tempo impacts negativily in popularity. □ □

Variable	Description	Type	Priority
popularity	value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity	Numeric al	0
year	Track year	Numeric al	0
genre	Track genre	Categori cal	0
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. A value of 0.0 is least danceable and 1.0 is most danceable.	Numeric al	0
energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.	Numeric al	0
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).	Numéric a	0
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.	Numeric al	0
time_signatu re	An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".	Numeric al	0
artist_name	The artists' names who performed the track. If there is more than one artist, they are separated by a ;	Categori c	1
loudness	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.	Numeric al	1
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.	Numeric al	1
instrumental ness	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". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.	Numeric al	1
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.	Numéric a	1
duration_ms	The track length in milliseconds	Numeric al	1
track_name	Name of the track	Categori c	2
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.	Numeric al	2
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.	Numeric al	2
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.	Numeric al	2

What this data set reveals?
Data cleaning

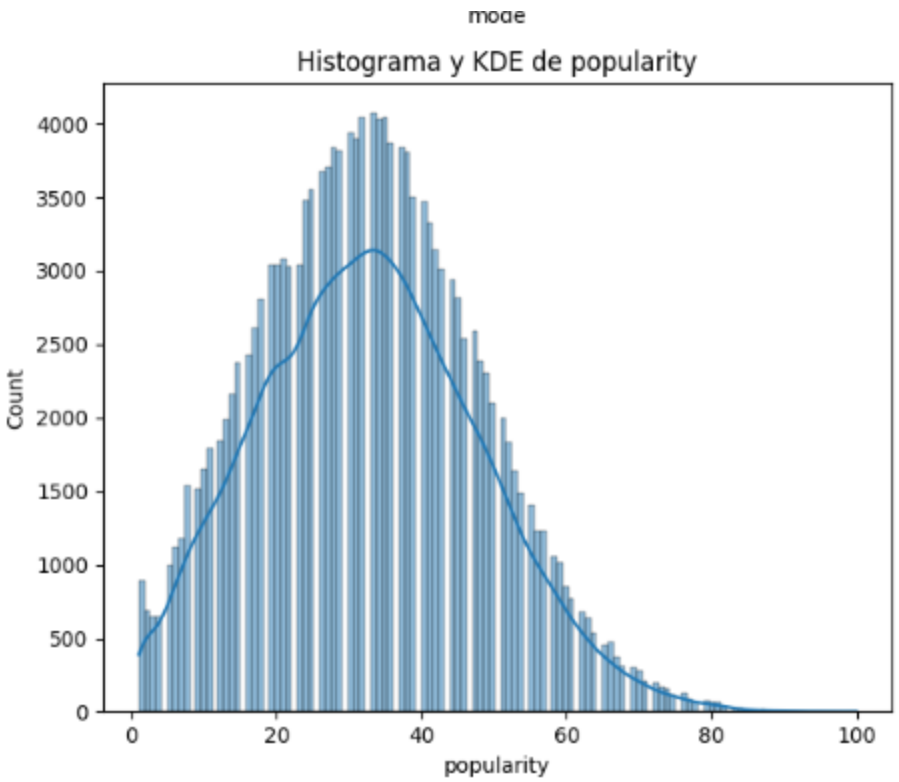
DATA CLEANING SCHEMA: from 1m instances to 200k

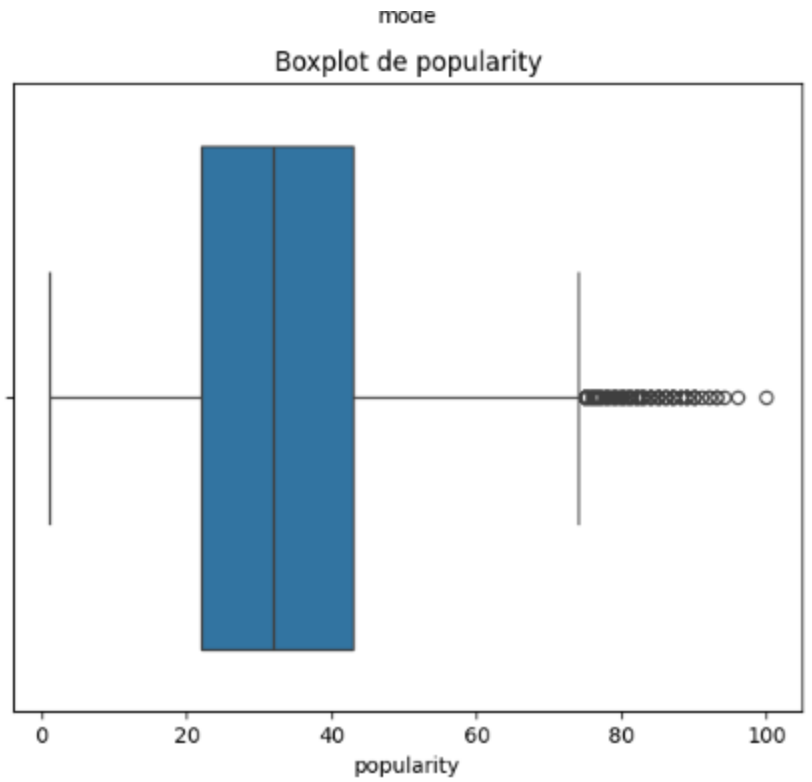


What this data set reveals?

Understanding Popularity

32 is a pop index mean, important level of outliers, relevant for this aim (1% of train data)



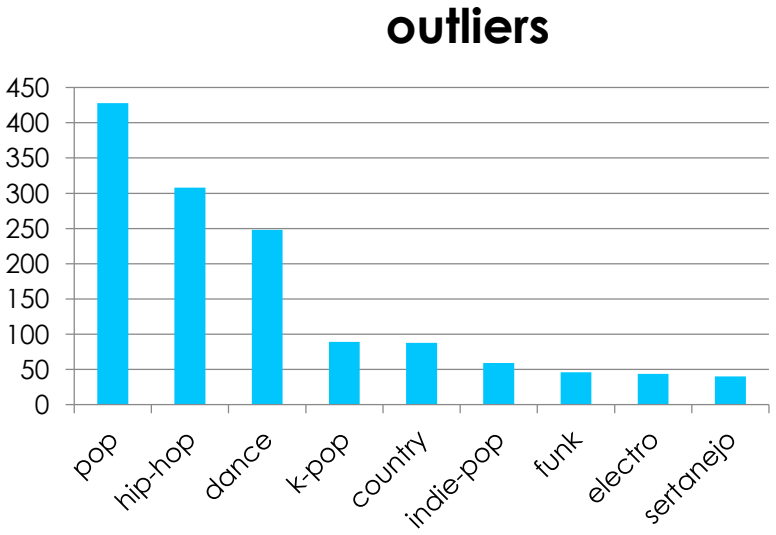


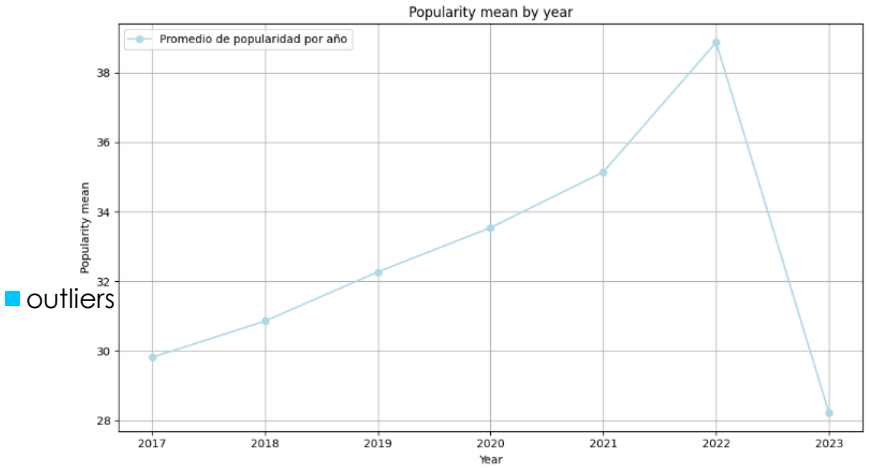
artist_name	popularity	year	genre
P!nk	81	2023	dance
The Chainsmokers	72	2018	dance
Jordan Davis	74	2022	country
keshi	82	2022	chill
Gusttavo Lima	71	2022	sertanejo
Måneskin	71	2023	indie-pop
Miley Cyrus	78	2020	pop
Morgan Wallen	80	2023	country
Billie Ellish	81	2019	electro
Mc Jacare	79	2022	funk

What this data set reveals?

Understanding Popularity

Pop or hip hop are genres more popular, also more representation.
This data set has a popularity mean decreasing in 2023

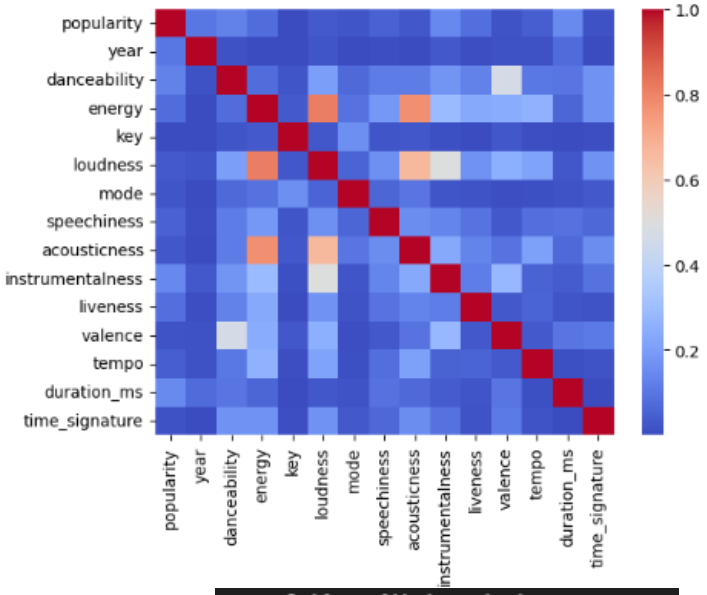




What this data set reveals?

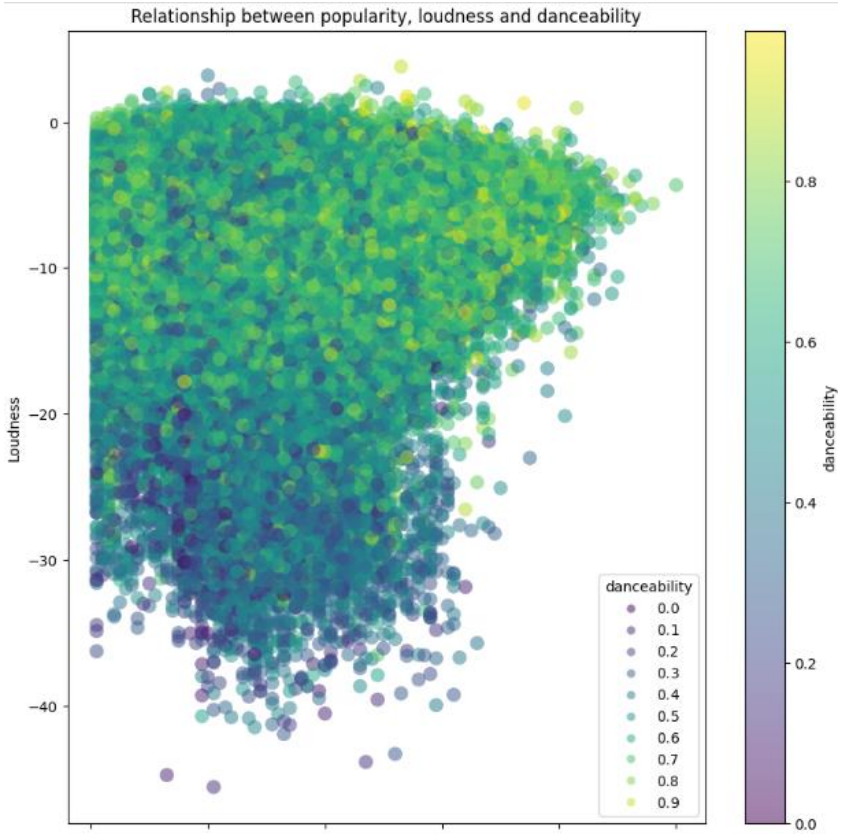
Understanding Popularity

Correlation with popularity is related to DURATION, INSTRUMENTAL, DANCING AND YEAR



Correlation with target above 0.03

duration_ms	0.145454
instrumentalness	0.141190
danceability	0.127397
year	0.098867
liveness	0.081357
energy	0.077282
speechiness	0.051018
tempo	0.040681
loudness	0.035239



Where words fail music speaks

Can machine learning help to accurate popularity?

(From a unexpert data science student work)

Machine learning
Key Issues

Features selection

Those correlating +0.03 with and without ARTIST

Those correlating +0.03 with and with GENRE

Features transformation

Categorial dummies or label encoder (machine power issue)

Numerical Logarithm or min max Scaler

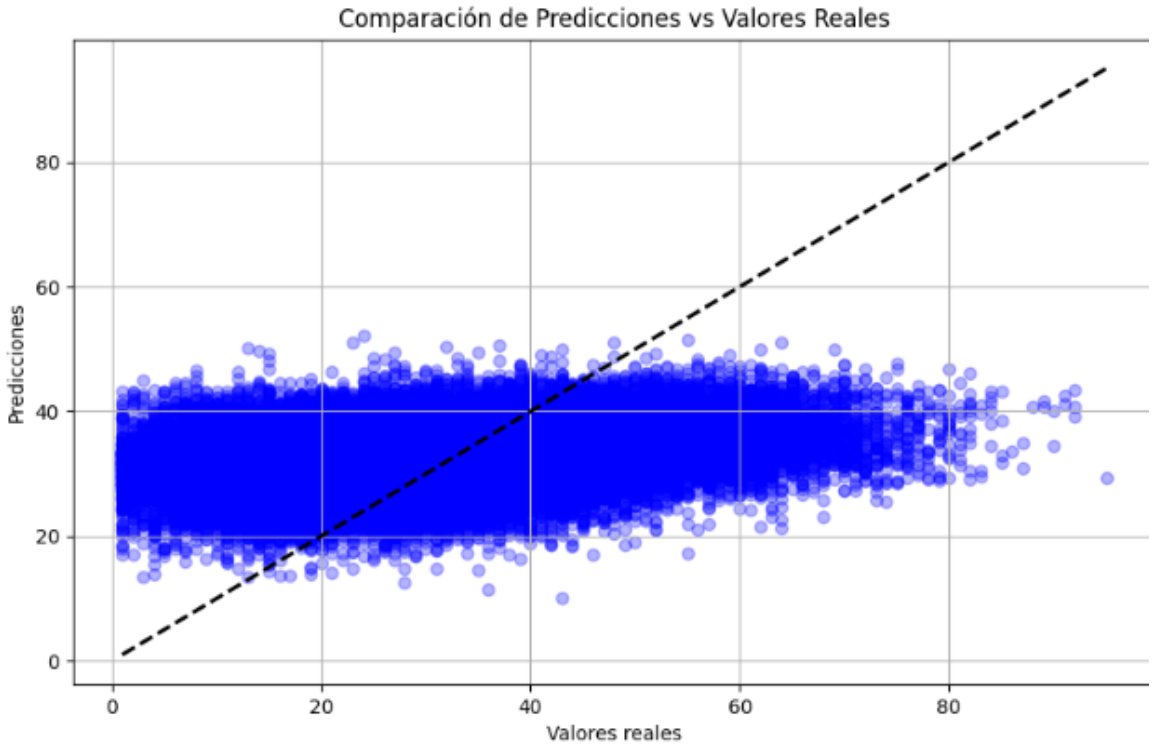
Modelling choosing

```
For SET minimum:
Lightgbm:  Mean MAE: -10.9491
XGBoost:  Mean MAE: -11.2472
Gradient Boosting:  Mean MAE: -10.9033
Linear Regression:  Mean MAE: -11.4980
******
For SET minimum_with_artist:
Lightgbm:  Mean MAE: -10.7723
XGBoost:  Mean MAE: -10.6038
Gradient Boosting:  Mean MAE: -10.1149
Linear Regression:  Mean MAE: -11.4949
******
For SET minimum_with_genre:
Lightgbm:  Mean MAE: -7.7261
XGBoost:  Mean MAE: -7.7868
Gradient Boosting:  Mean MAE: -7.5338
Linear Regression:  Mean MAE: -11.4904
******
```

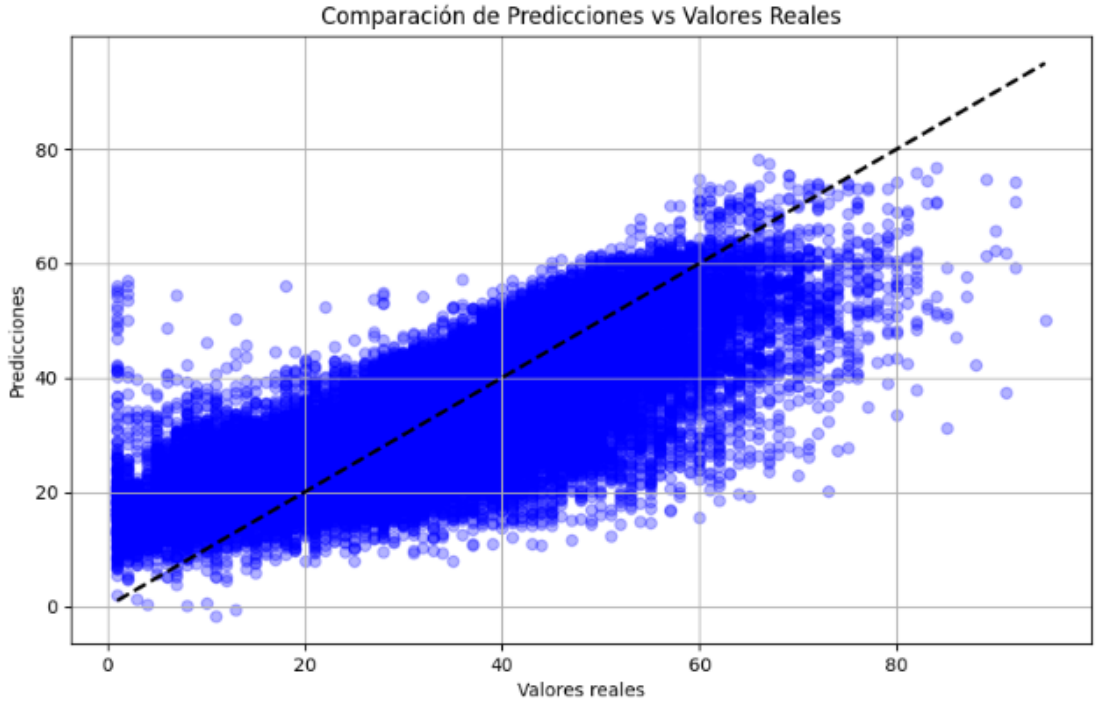
Machine learning
Fit models

Tasas de reciclado al alza en ritmos diferentes según la zona.

Linear Regression (MAPE 0,79)



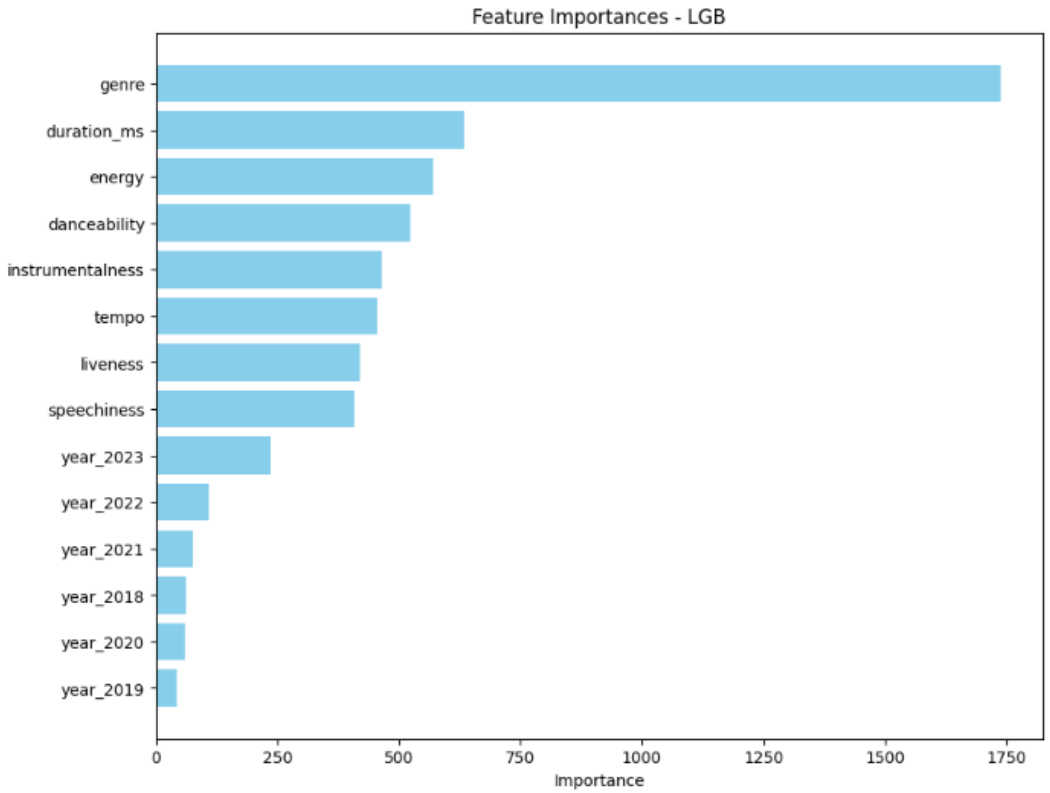
Light GBM (MAPE 0,47)



Machine learning
Fit models

Feature importances of the best model

Light GBM




Where words fail music speaks
Predictions

Predictions
Issue example

This model and project provides a prediction instrument, in order to add new inputs and predict the popularity

```
# Input
new_input = {
    'artist_name': 5965,
    'genre': 21,
    'danceability': 0.49,
    'energy': 0.304,
    'key': 2,
    'loudness': 0.7284403298016733,
    'mode': 0,
    'speechiness': 0.0515,
    'acousticness': 0.836,
    'instrumentalness': 0.912,
    'liveness': 0.0923,
    'valence': 0.343,
    'tempo': 0.47874244842705027,
    'duration_ms': 11.845640652713728,
    'time_signature': 0.8,
    'year_2017': 1,
    'year_2018': 0,
    'year_2019': 0,
    'year_2020': 0,
    'year_2021': 0,
    'year_2022': 0,
    'year_2023': 0
}
```

```
[ ]  new_prediction_lgb = best_lgb_reg_short.predict(new_input_df[selected_features])

     print("Nueva predicción:", new_prediction_lgb)
```

```
Nueva predicción: [15.27421991]
```

In conclusion, strong marketing factors impact on popularity, others can be extracted from this project:

If you want the most popular song, choose genre, dancing or instrumental factors.

Nevertheless each genre has its popularity range

Artist was relevant but not so far

This project can be improve with a hyperparamenter stronger evaluation and largest data sets

Thanks!

Appendix

Insights

Data set

https://www.kaggle.com/datasets/ziriantahirli/million-song-data-analysis-2/data

- Github/MariaRepository