Spotify popularity based on virality, nothing else matters # Metallica?

Popularity Index study, feature importances and prediction

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

What this presentation is about?

1. Brief Spotify context
2. What data shows?
3. Machine Learning speaking: Can help to accurate popularity?
4. Predictions

Music accompanies in our best and worst moment –
Daniel Ek (Spotify CEO)

Brief Spotify context

Spotify and music context

Spotify figures

- €13.24 billion revenue 2023 (+12.9% vs LY)
- 551 million month users
- 220 million suscriptors
- More than 100 million songs

Spotify and music context

Songs's spotify popularity is related to several factors, artist and producers are interested in revealing **formula**. In Spotify words, Popularity Index:

- Is a **dynamic** metric
- Influenced by played in **recent period** (usually 30 days).
- **Listener interaction** affects the score.
- **Freshness of streams**
- Based on **global** plays, local popularity can also impact.

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

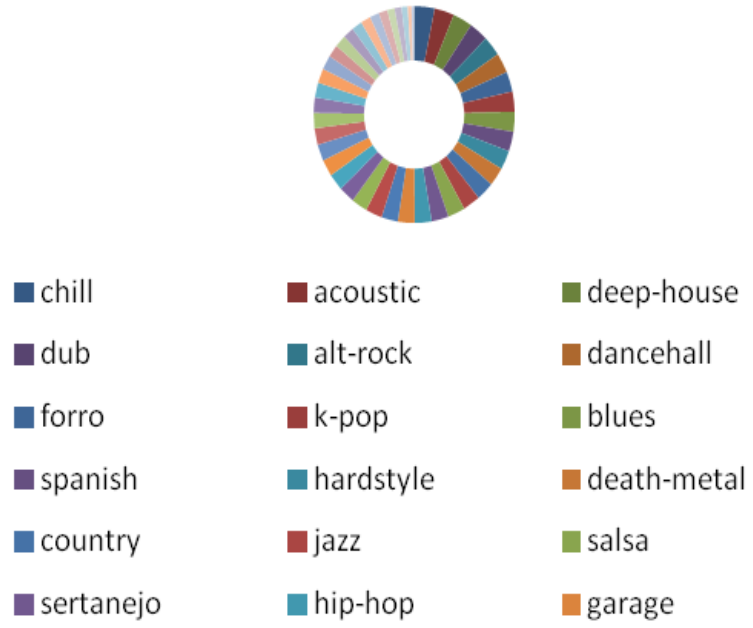
Music is an echo of the invisible world - Manzini

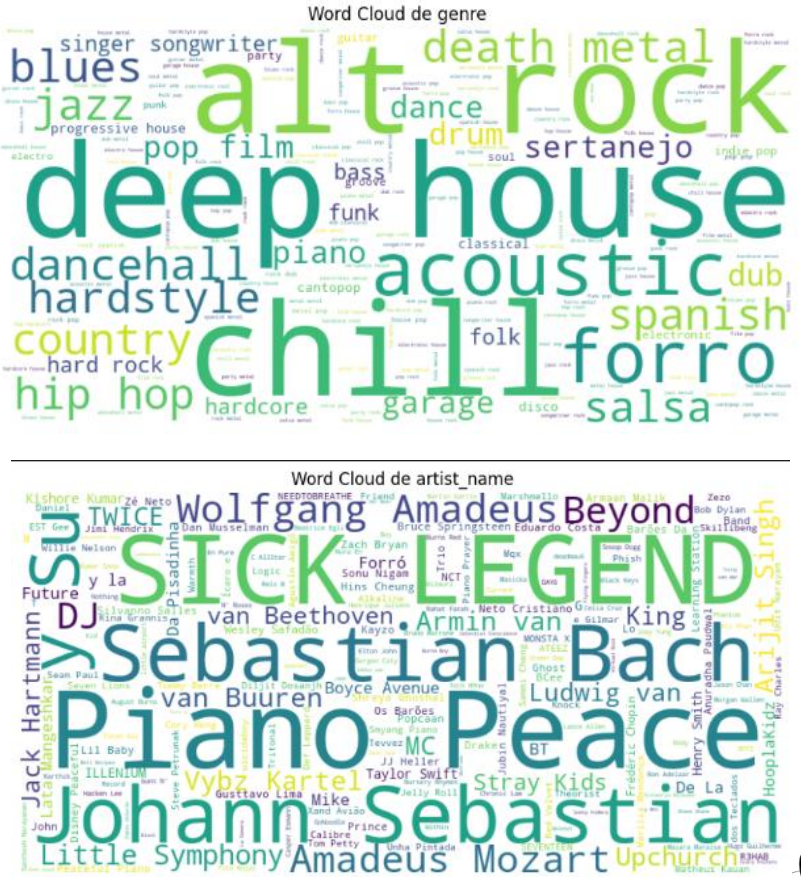
What this data set reveals?

What this data set reveals?

Plenty of data

An ocean of music and features to understand and trends from 2017-2023.





What this data set reveals?
Features & Target

****Track type****
A. Year
B. Genre
C. Artist_name
D. Track_name

****Track composition****
A. Loudness
B. Acousticness
C. Instrumentalness
D. Liveness
E. Speechiness
F. Key
G. Mode

****Track rythm****
A. Danceability
B. Energy
C. Valence
D. Tempo

****Track timings****
A. Time_signature
B. Duration_ms

What this data set reveals?
Features & Target

□ □ 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 on popularity □ □

What this data set reveals?
Features & Target

DATA CLEANING

SCHEMA:

from 1m instances

to 200k

- Period previous to 2017 no to be considered
- Columns to drop: track_id, track_name
- Genres irrelevant to drop
- 0 popularity tracks to drop

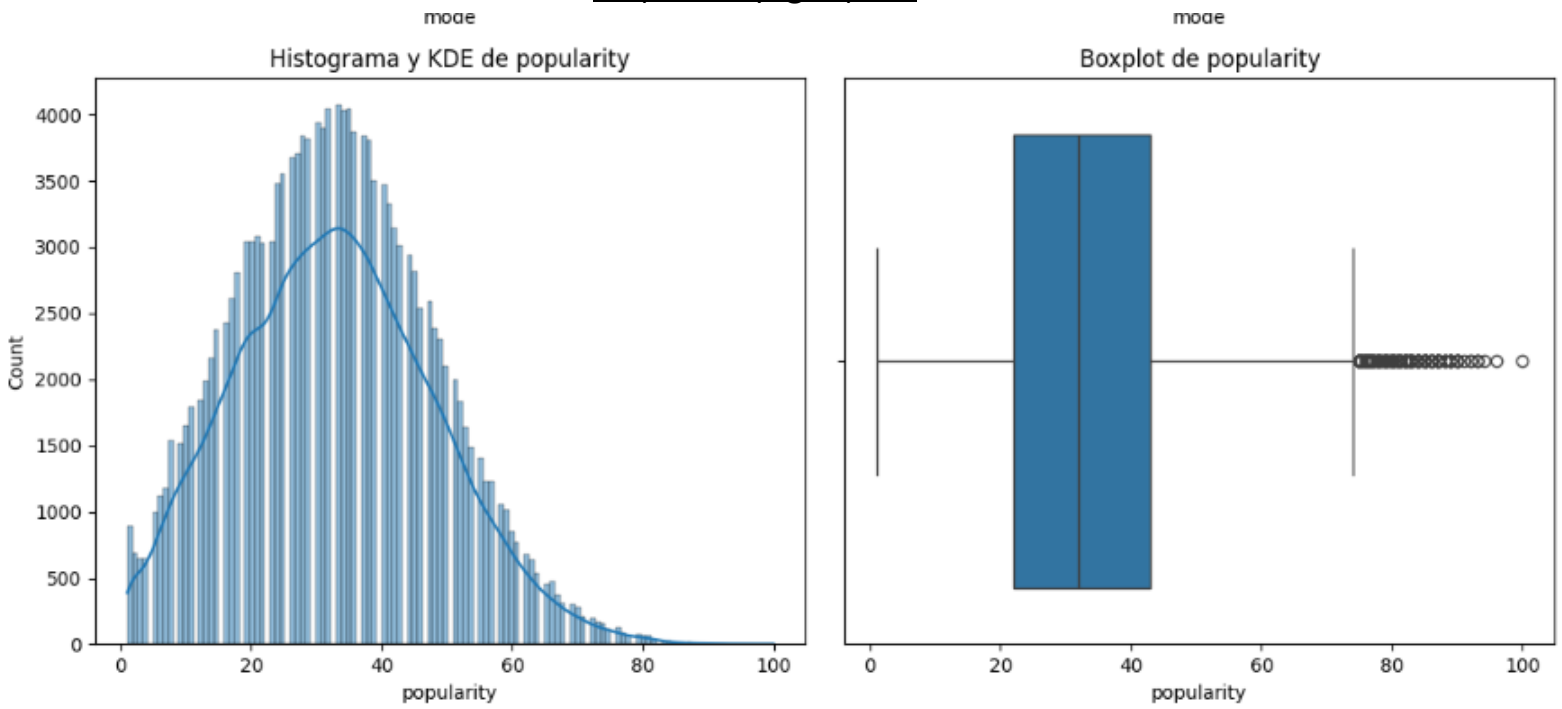
What this data set reveals?

Features & Target: *Understanding Popularity*

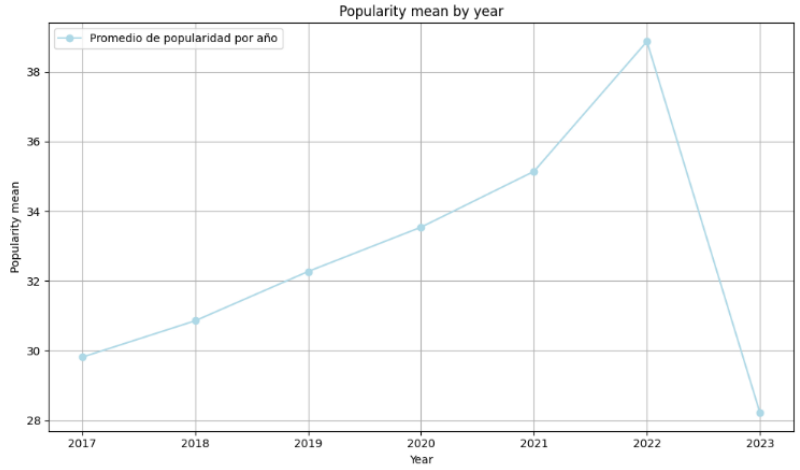
Mean popularity is 32. There are important level of outliers (1% of train data), are KEY for this case

This data set has a popularity mean decreasing in 2023

Popularity graphs



Popularity evolution



What this data set reveals?

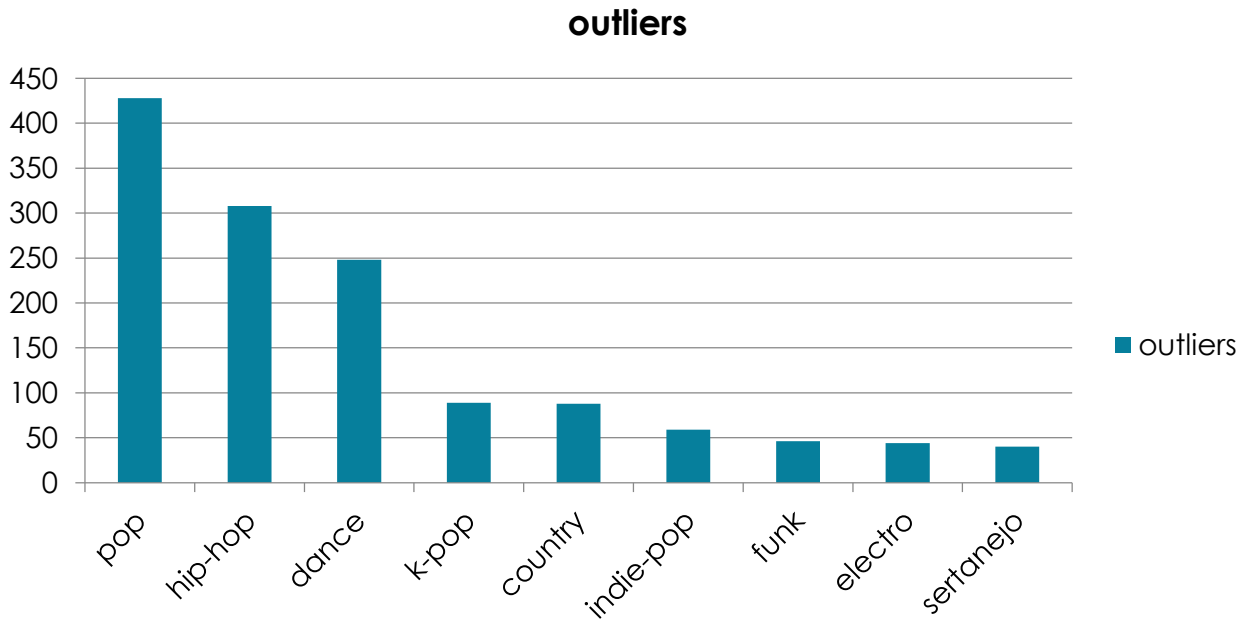
Features & Target: *Understanding Popularity*

Pop or hip hop are genres more popular, also more representation.

Most popular tracks

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

Top genre popular tracks



What this data set reveals?

Features & Target: *Understanding Popularity*

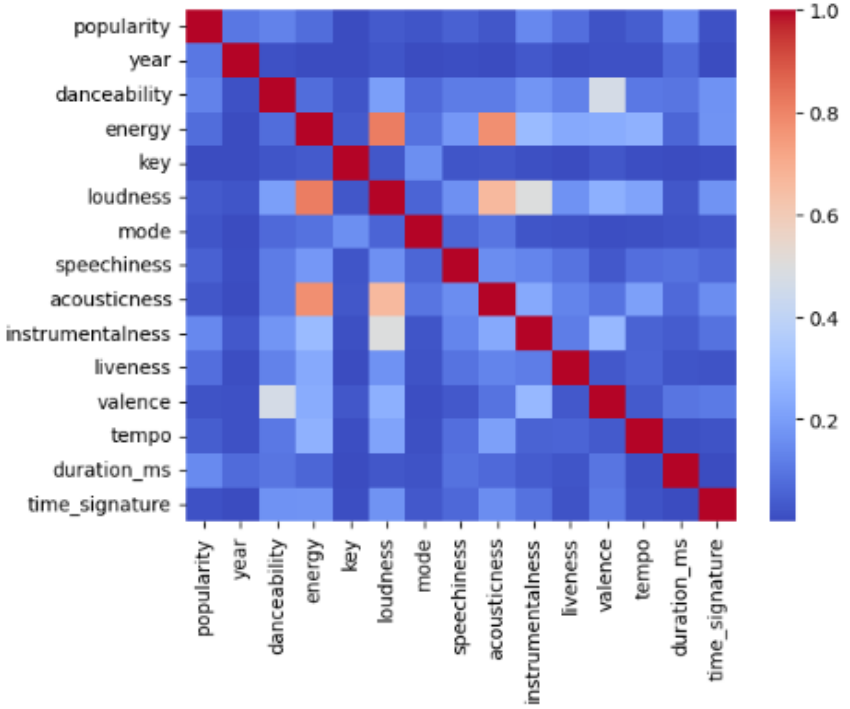
Correlation with popularity is related to **DURATION, INSTRUMENTAL, DANCING AND YEAR**

Features correlation with popularity

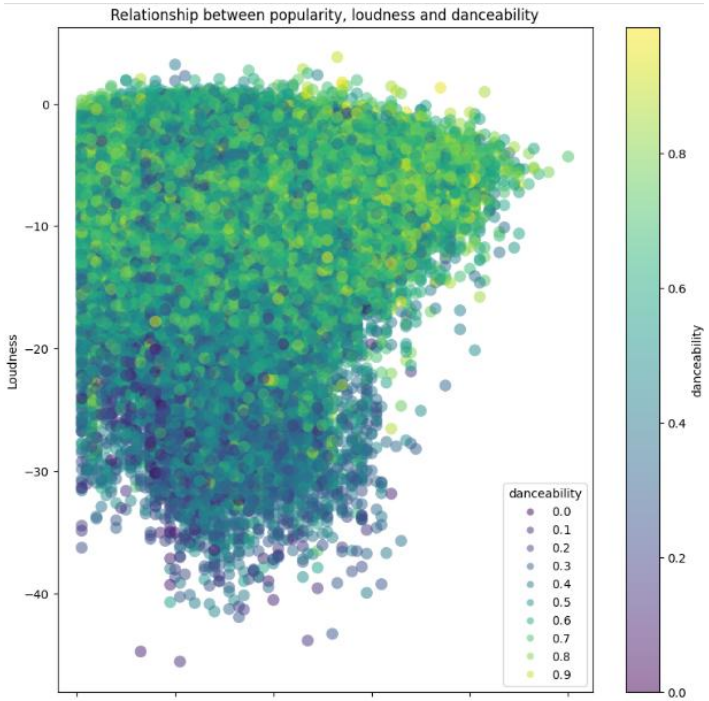
```
 Correlation with target above 0.03

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
```

Features colinearity



Tridimensional analysis



Where words fail music speaks -Hans Christian Andersen

Machine Learning speaking: Can help to accurate popularity?

(From a unexpert data science student work)

ML speaking: Can help to accurate popularity?

The band:

Features selection

- Features correlating +0.03 with and without ARTIST
- Features correlating +0.03 with GENRE

Features transformation

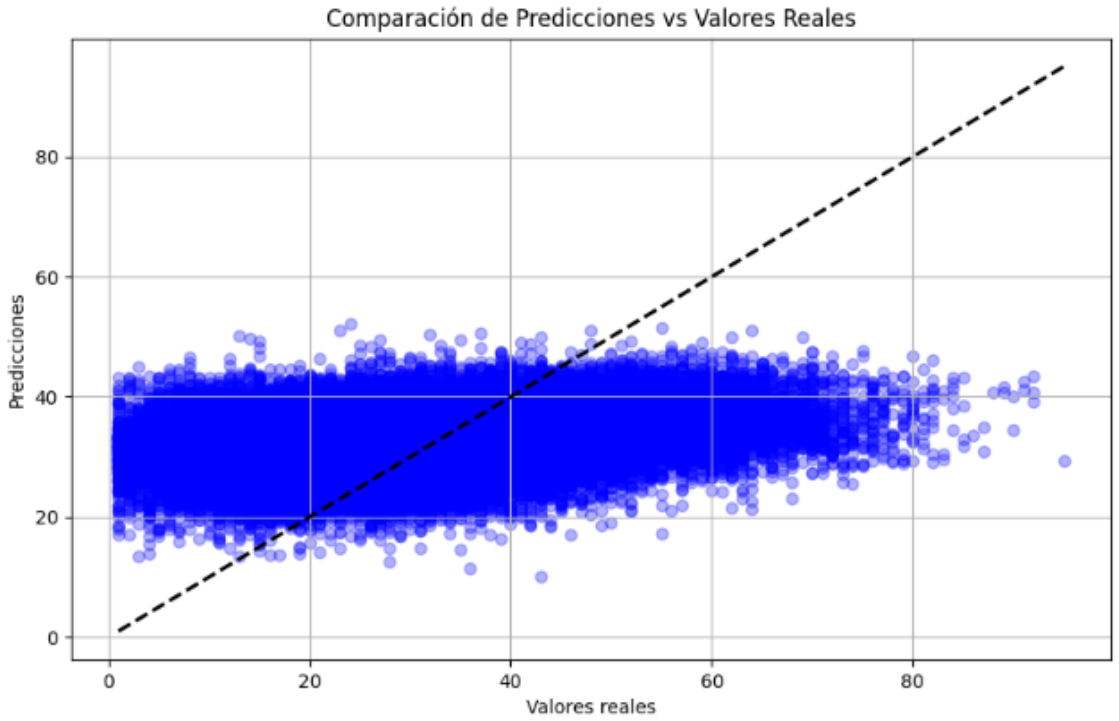
- Categorial dummies or label encoder (machine power issue)
- Numerical Logarithm or min max Scaler

Modelling choosing

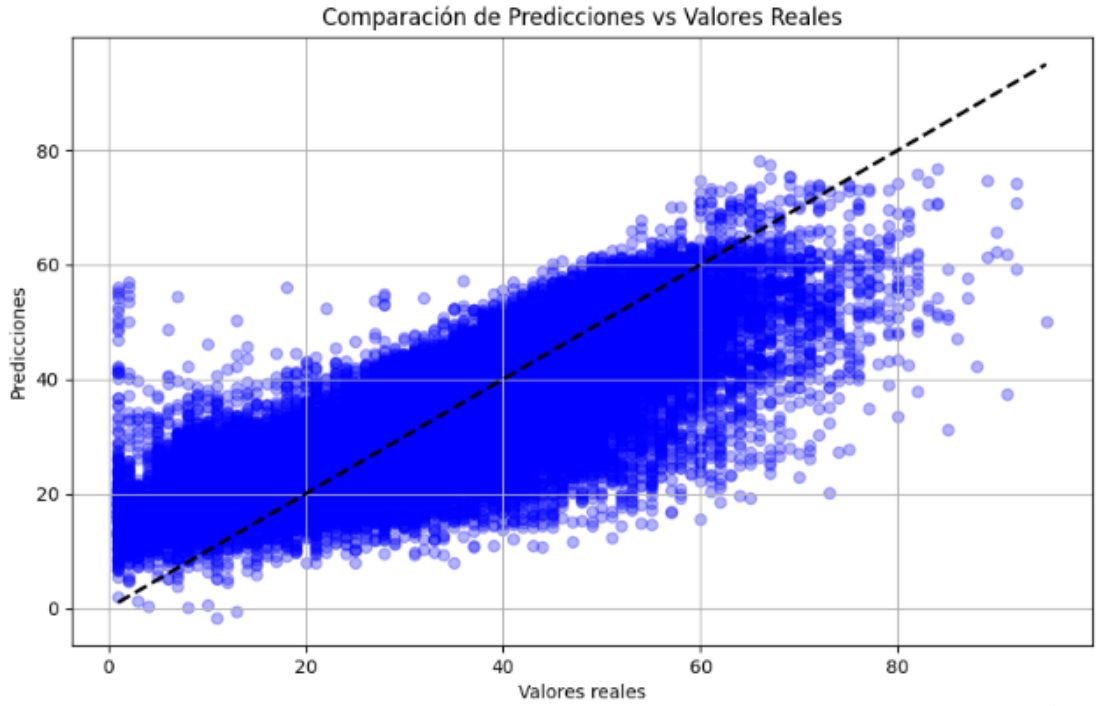
```
For SET minimum:
Lightgbm:  Mean MAE: -10.9491
XGBoost:  Mean MAE: -11.2472
Gradient Boosting:  Mean MAE: -10.9033
Linear Regression:  Mean MAE: -11.4980
******
For SET minimum_with_artist:
Lightgbm:  Mean MAE: -10.7723
XGBoost:  Mean MAE: -10.6038
Gradient Boosting:  Mean MAE: -10.1149
Linear Regression:  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
******
```

ML speaking: Can help to accurate popularity?
Model matters

Linear Regression (MAPE 0,79)



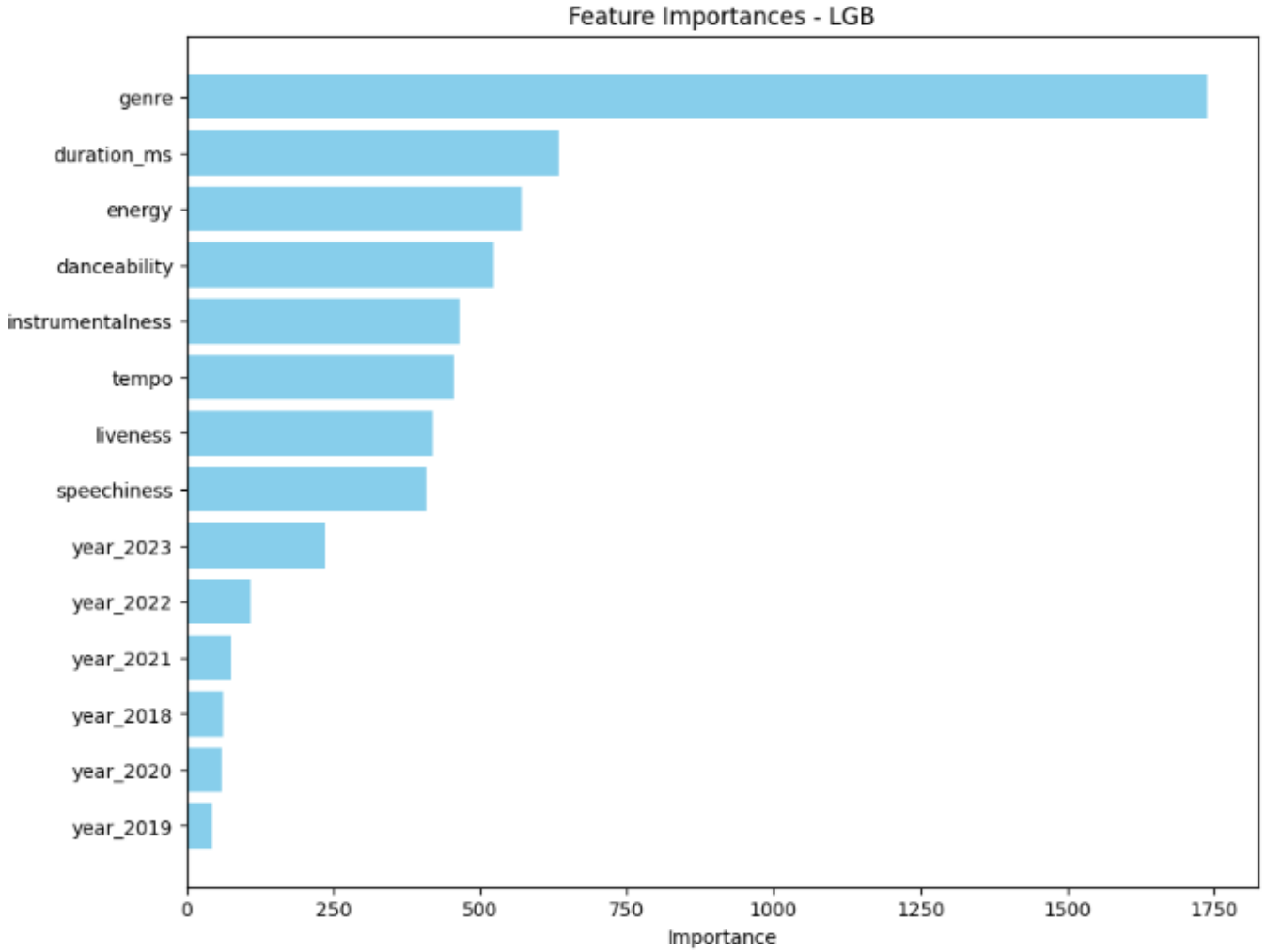
Light GBM (MAPE 0,47)



ML speaking: Can help to accurate popularity?

Model matters

Feature importances of the best model Light GBM



ML speaking: Can help to accurate popularity?

Model matters

A lot may be improved

Data model seems a Babel Tower

What and how?

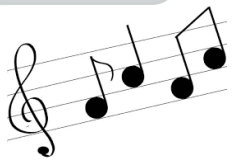
ML speaking: Can help to accurate popularity?

Model matters

Let´s just predict chill?!

Modeling data sets clustered by **GENRE** shows different performance

Chill	• Including feature artist improves MAE 0.2 • Linear Reg MAPE 0.14; LightGBM 0.14
Acoustic	• Including feature artist worsen MAE 0.2 • Linear Reg MAPE 0.57; LightGBM 0.5
Spanish	• Including feature artist improves MAE 0 • Linear Reg MAPE 0.42; LightGBM 0.39
Pop	• Including feature artist improves MAE 0.8 • Linear Reg MAPE 0.26; LightGBM 0.28



ML speaking: Can help to accurate popularity?
Model matters

Modeling data sets clustered by **GENRE** shows different performance

Different feature importance scores

Future approachs needed:

model by genre, increasing data in some genres, non supervised models…

Music is the shorthand of emotion – León Tolstói

Popularity Predictions

Popularity predictions
Prototype

This project provides a prototype prediction tool:

A new input song can be introduced

```
# 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
}
```

The model will predict propularity score

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

     print("Nueva predicción:", new_prediction_lgb)

     Nueva predicción: [15.27421991]
```

Popularity predictions
Prototype

Wanna evaluate your composition's popularity 💔?

Your composition is great as it is, but let's see what Spotify may score.

Select your genre:

pop

Spotify babel Tower made with IA

Enter values to evaluate your composition:

Select duration in ms (10-15):
13
10 15

Select instrumentalness (0-1):
0.42
0.00 1.00

Select danceability (0-1):
0.81
0.00 1.00

Select liveness (0-1):
0.65
0.00 1.00

Select energy (0-1):
0.44
0.00 1.00

Select speechiness (0-1):
0.33
0.00 1.00

Select tempo (0-1):
0.34
0.00 1.00

Popularity predictions
Prototype

Select the year of the composition:

2022

Confirm All

Prediction: 57.54

Trust yourself, we are more than a NUMBER

Wanna learn more about composition? Check our courses at

www.compositionmaster.com

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.
- Each genre has its popularity range
- Artist was relevant but not so far
- Model prototype to be improved: hyperparamenter stronger evaluation, largest data sets & non supervised models

Thanks!

Appendix

Insights

Data set

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

- Github/MariaRepository