Building a predictive model for songs ranking in top 5

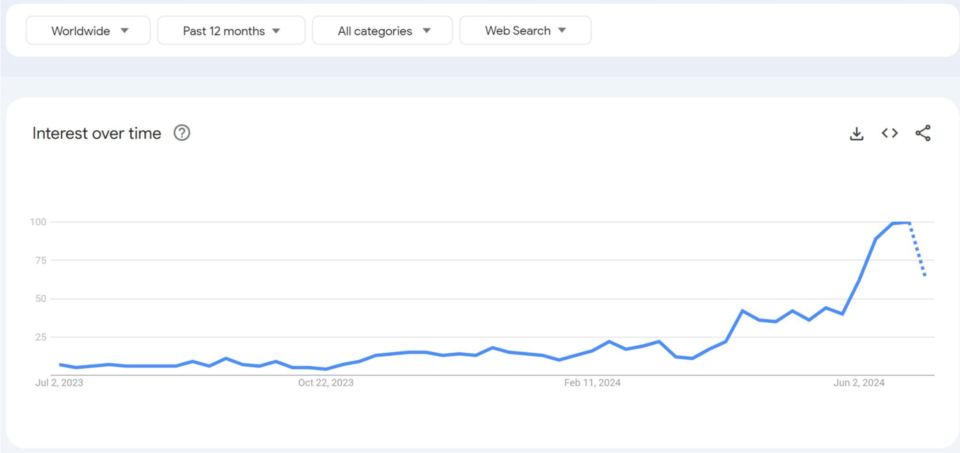
Alfredo Hernandez, July 2024

Springboard

Problem statement

- >200M songs worldwide
  - >3M new songs/year
- 73% people use streaming services
  - 0.5 ¢ / song played

How can we predict which songs will be in the top 5, so we can maximize the revenue coming from streaming services?



Music

Data Sources

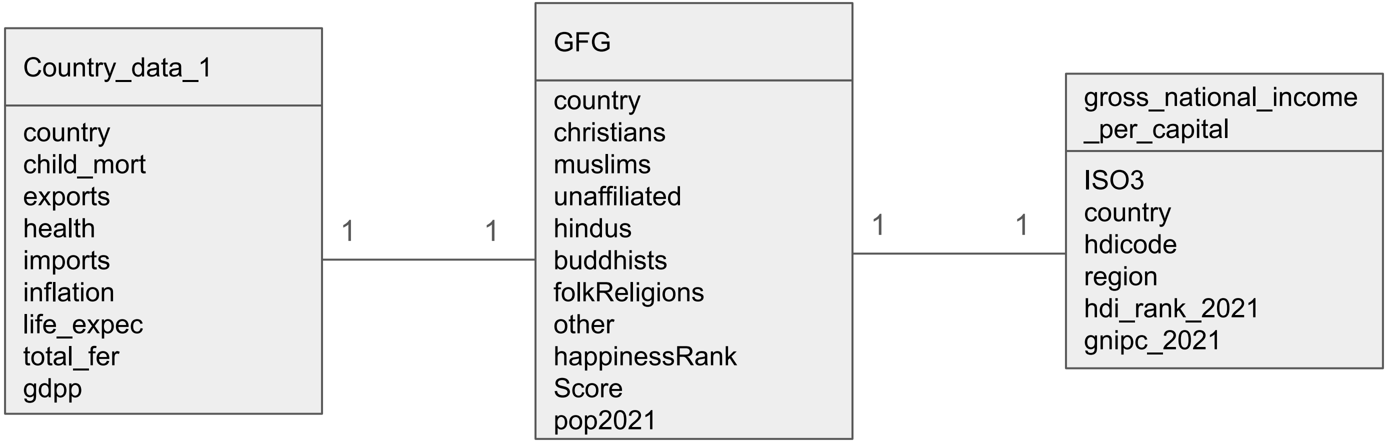
- Top Spotify Songs in 73 Countries (Daily Updated)
- Countries by population happiness index & religion
- Global Socio-Economic & Environmental Indicators
- Socio-Economic Spread of Countries

kaggle

Songs Data Overview

- Song rank position on given date at a given country
  - Song features repeat across all days/countries
  - Rank and movement change with country and date

Socio-Economic Data Overview

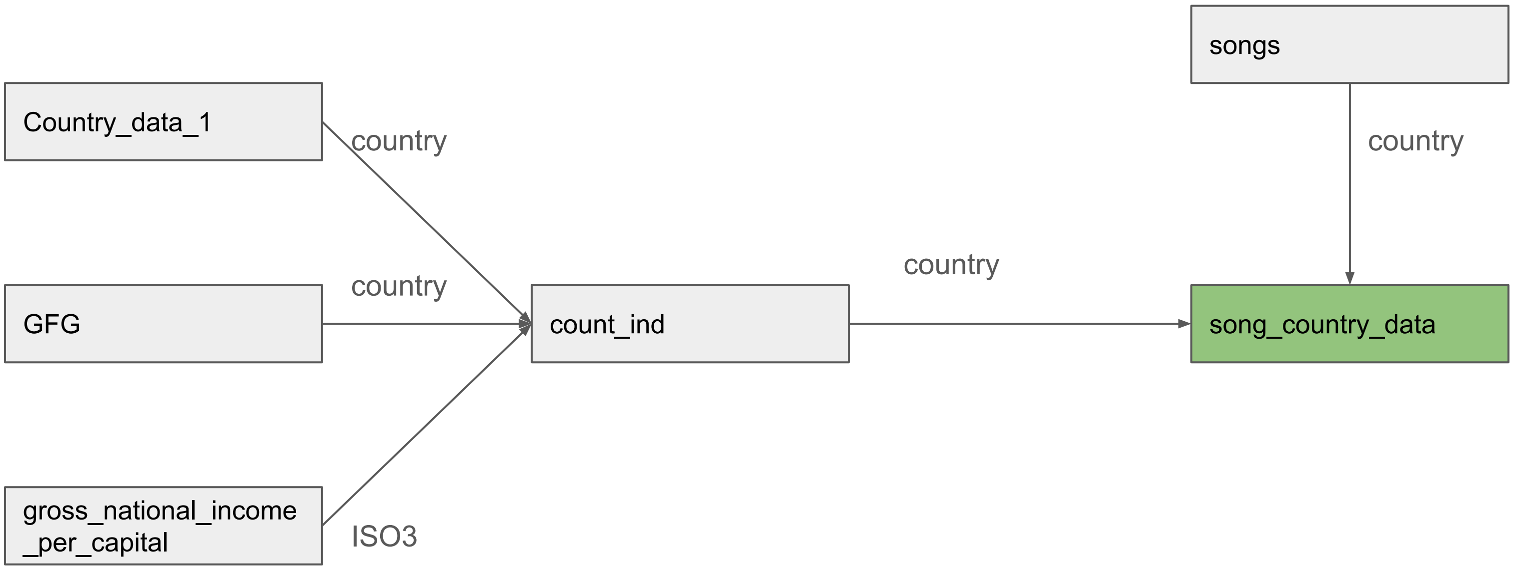


Socio-Economic Spread of Countries

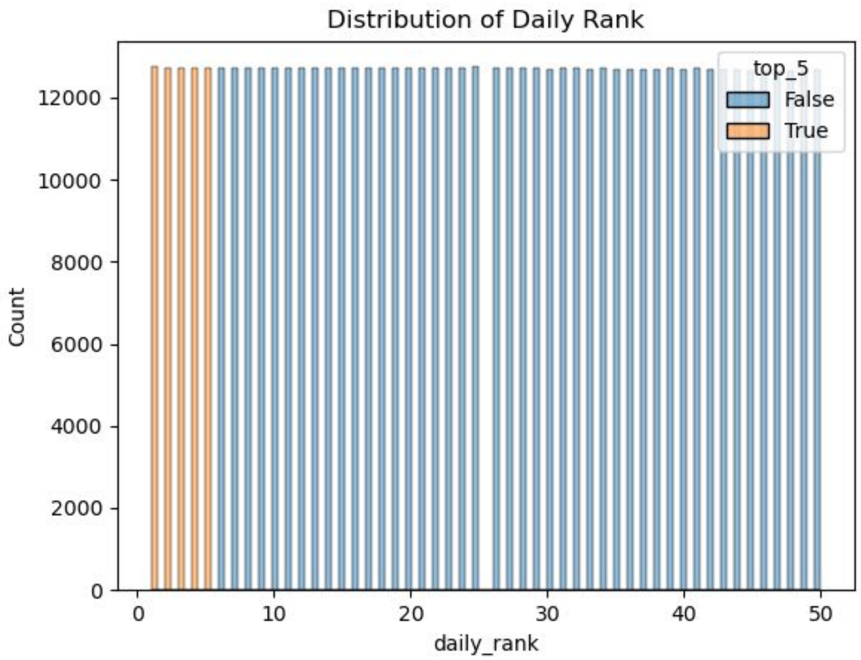
Countries by population happiness index & religion

Global Socio-Economic & Environmental Indicators

Data Merging

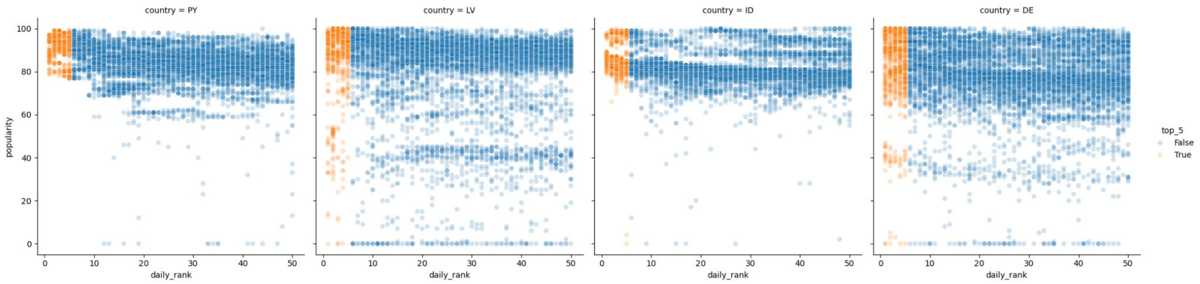


Exploratory Data Analysis (EDA)

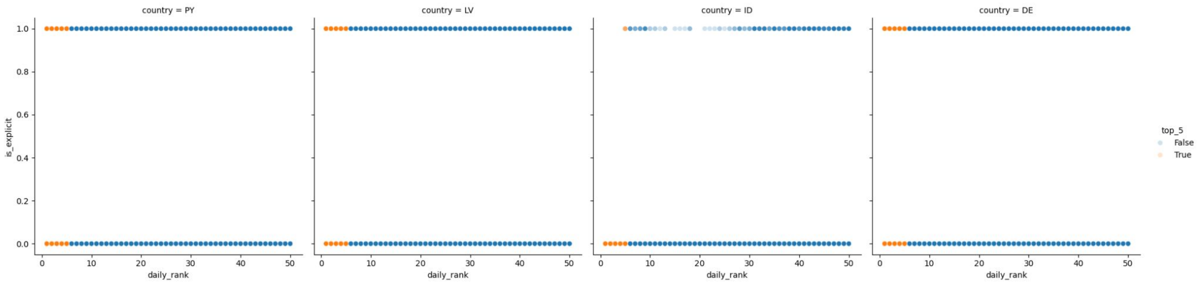


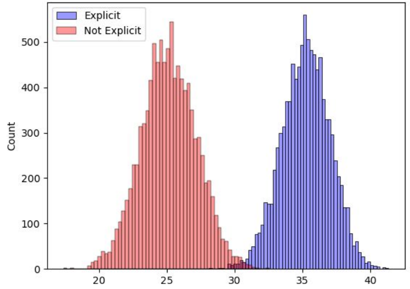
- Daily_rank uniform distribution
- High correlation:
  - Loudness - energy
  - Imports - exports
  - Income - gnipc and gdpp
  - Hdi_rank_2021 - several features.

Exploratory Data Analysis (EDA)

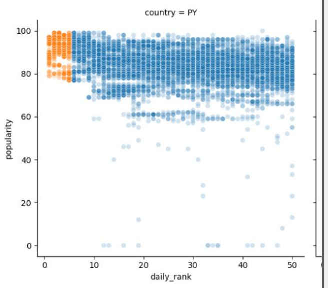


- Latvia and Germany have songs with lower popularity in top5
- Indonesia rarely places explicit songs on top5.
  - Explicit songs place 35
  - Non explicit place 25.



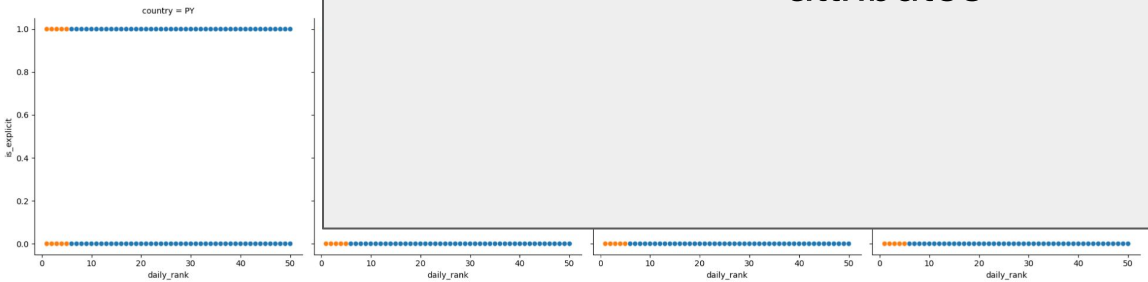


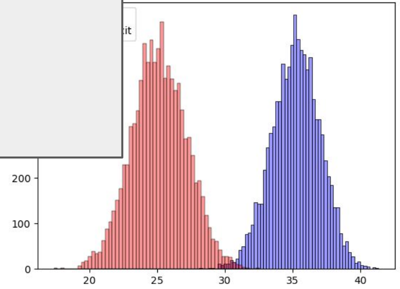
Exploratory Data Analysis (EDA)



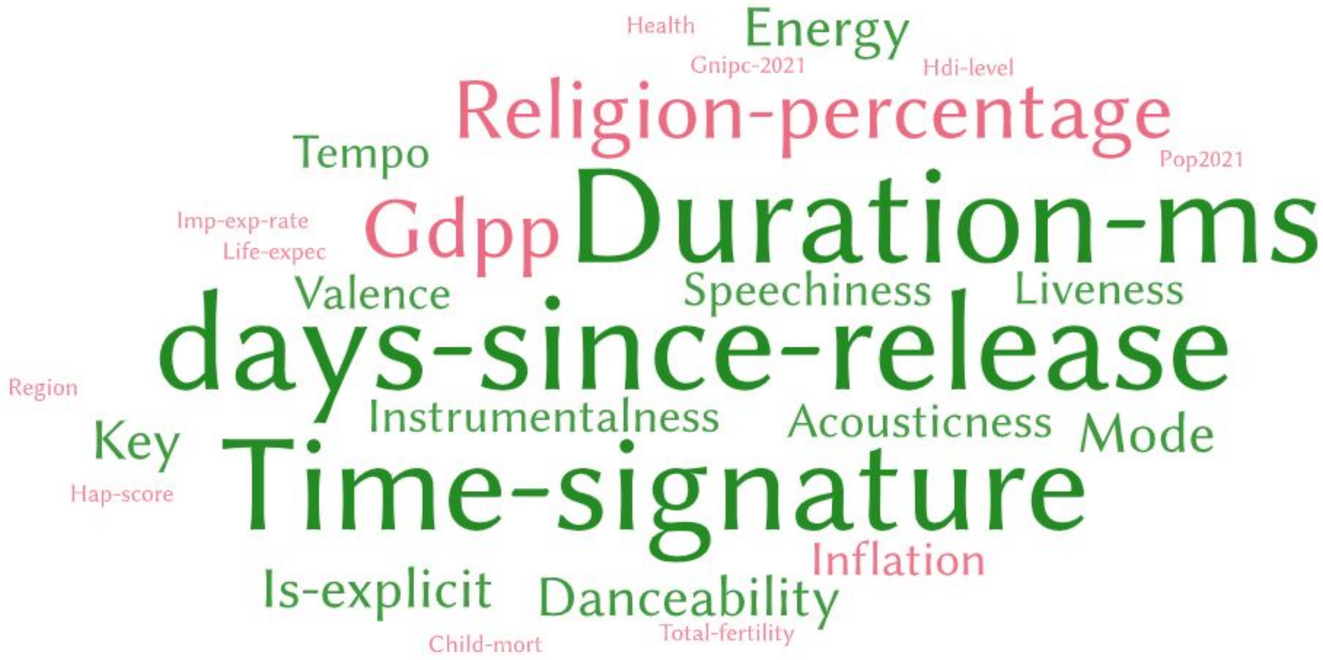
There is some evidence to conclude that top songs are influenced by the country as well as the songs attributes

and Germany
songs with lower
arity in top5
esia rarely places
t songs on top5.
Explicit songs place 35
Non explicit place 25.





Feature Selection



Model Selection & metrics

- Constant
- KNN Classifier
- Logistic Regression
- Random Forest
- Gradient Boosting

- **F1**
  - Imbalanced data (10% True)
- Balanced Accuracy
- Recall

scikit learn

Modeling

Data Pre-processing:

1. Feature Selection
2. Label Encoding
3. Data Split train/test sets (30% test). Maintaining same positive rate
4. Scaling

Modeling and testing:

1. Select model to test
2. Fit the model using default hyperparameters
3. Predict test results
4. Calculate metrics

For every model

Selecting the best model

Gridsearch hyperparameter tunning

1. 5 fold CV
2. Max_depth 10 steps between 1 - 100

Starting from the results previously found

Bayesian Optimization

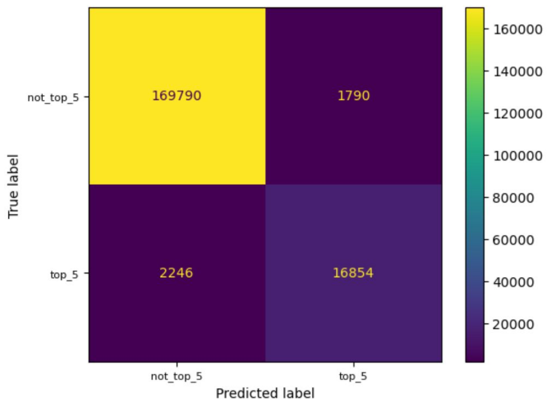
1. 3 fold CV
2. Using F1 for scoring
3. Optimized for max_depth, min_samples_split, min_samples_leaf, n_estimators.
4. Used 5 iter, 2 init_points.

Best Random Forest Model

max_depth = 66
min_samples_split = 2
min_samples_leaf = 2
n_estimators = 129

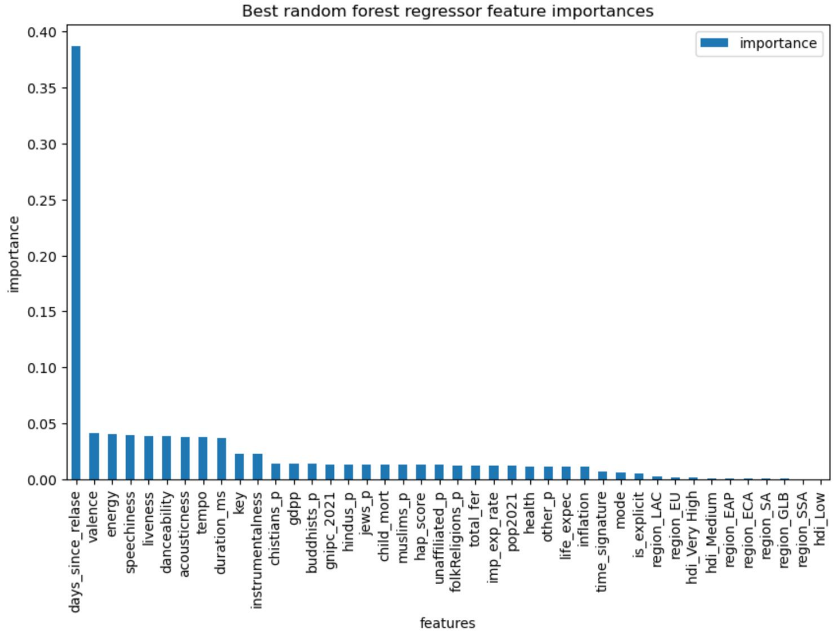
Results

	Balanced Accuracy	Precision	Recall	F1
Constant	50.0%	0.0%	0.0%	0.000
KNN Classifier	93.1%	88.5%	87.5%	0.880
Logistic Regression	55.5%	12.3%	53.1%	0.200
Random Forest	93.6%	90.0%	88.2%	0.891
Gradient Boosting	51.6%	72.6%	3.4%	0.064
Random Forest (optimized)	93.6%	90.4%	88.2%	0.893



Top Features

- Days_since_release -> 40%
- Song attributes -> 45%
- Socioeconomic indicators ->15%
  - Religion
  - GDP



Conclusions and next steps

There is potential to predict whether a song will reach the top 5 by analyzing song characteristics and socio-economic indicators. Our results show that a combination of both musical attributes and socio-economic factors can significantly enhance the accuracy of these predictions.

To build on the current findings, future work should focus on the following areas:

- **Country-Level Analysis:**
  - Dive deeper into modeling the data at a country-specific level.
  - Concentrate on countries with the highest number of listeners or those providing the best return on investment (ROI) for the music industry.
- **Temporal Dynamics:**
  - Analyze temporal trends to understand how song popularity evolves over time.
  - Incorporate time series analysis to predict future trends and identify early indicators of potential hits.
- **Genre-Specific Models:**
  - Develop separate models for different music genres to account for genre-specific characteristics and listener preferences.
  - Compare the performance and predictive power of genre-specific models versus a general model.
- **Enhanced Feature Engineering:**
  - Explore additional features such as social media presence, marketing efforts, and cultural factors.
  - Incorporate advanced natural language processing (NLP) techniques to analyze song lyrics for thematic and emotional content.
- **Model Optimization and Validation:**
  - Experiment with more sophisticated machine learning algorithms and ensemble methods to improve prediction accuracy.

By pursuing these next steps, we can further refine our understanding of the factors driving song popularity and enhance our predictive capabilities. This deeper insight will provide valuable guidance for artists, producers, and marketers aiming to maximize the reach and impact of their music in the global market.

Q&A