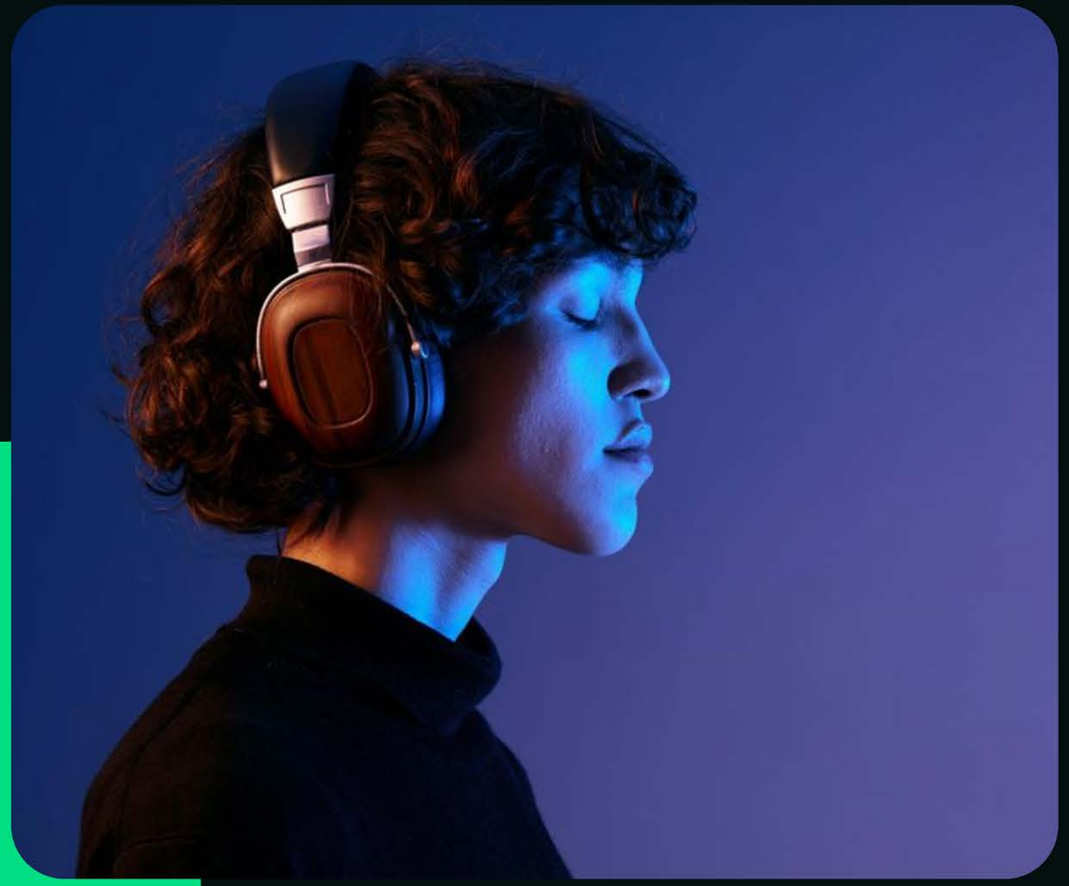
Spotify Insights Presentation

Project 4 - Group 6
Adam Loux
Misha Mambully Muralidharan
Cecilia Rocha
Willian Ruiz

Inspiration & Purpose



Music is a universal language that resonates with everyone, and in the age of data, we can explore it in entirely new ways.

Data & Data Cleaning

Let's Get Started

Yes, You Can!

Data Preparation for Machine Learning: Unnecessary columns were removed, and missing values in numerical columns were imputed using the median with SimpleImputer. The target variable, 'danceability', was binned into low, medium, and high for classification.

Data Cleaning for Tableau: Columns were grouped for clarity, and missing values were filled to ensure consistency for visualization purposes.

Research Questions

Can we classify songs into danceability categories (Low, Medium, High) based on audio features? 01

What mood categories correlate with the highest streams and the most popular songs? 02

How do audio features like energy, tempo, loudness, and danceability vary across music genres? 03

Data Analysis

You've Got It!

Machine Learning

The goal of the machine learning model was to classify songs based on their danceability, categorizing them into low, medium or high. Initially, we attempted to build a regression model to predict the popularity of songs using various audio features. However, the R-squared values for all the models were very low, indicating that the regression approach was not effective for our dataset. So we decided to shift the focus to a classification model, with the target variable being the danceability of the song.

You've Got It!

Machine Learning

We experimented with multiple machine learning algorithms such as Random Forest, Logistic Regression, and Gradient Boosting. After evaluating the models, we found LightGBM to be the most effective in classifying danceability.

You've Got It!

Machine Learning

```
TRAIN METRICS
    Confusion Matrix:
    [[19887    46  7680]
 [   99  4626  3936]
 [ 5474   985 42767]]

    AUC: 0.9127782735566777

    Classification Report:
                  precision    recall  f1-score   support

            High       0.78      0.72      0.75     27613
             Low       0.82      0.53      0.65      8661
          Medium       0.79      0.87      0.83     49226

        accuracy                           0.79     85500
       macro avg       0.80      0.71      0.74     85500
    weighted avg       0.79      0.79      0.78     85500
```

You've Got It!

Machine Learning

```
TEST METRICS
    Confusion Matrix:
    [[ 6366    24  2759]
 [    37  1481  1468]
 [ 1992   405 13968]]

    AUC: 0.8942835902806459

    Classification Report:
                  precision    recall  f1-score   support

            High       0.76      0.70      0.73      9149
             Low       0.78      0.50      0.60      2986
          Medium       0.77      0.85      0.81     16365

        accuracy                           0.77     28500
       macro avg       0.77      0.68      0.71     28500
    weighted avg       0.77      0.77      0.76     28500
```

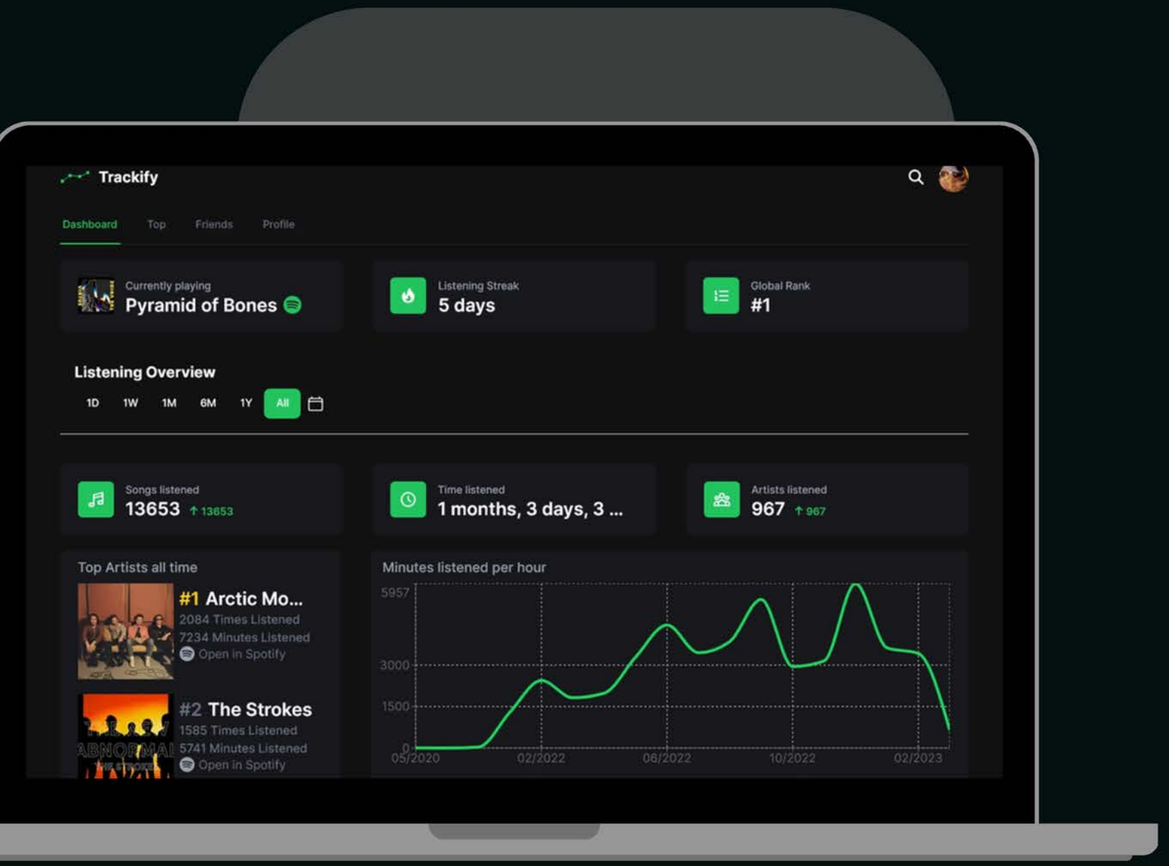
You've Got It!

Machine Learning

You've Got It!

We also wanted to implement a recommender system to suggest tracks to users. To do this, we used a content-based approach, where recommendations are made based on the features of the tracks, rather than on user behavior or preferences. In our case, we used the K-Nearest Neighbors (KNN) algorithm to recommend tracks based on their features such as popularity, danceability, energy, loudness, speechiness, and acousticness etc.

Content-based filtering focuses on the features of the items. Here, instead of relying on user behavior data like ratings or play history we recommend tracks that are similar in terms of their features.



Spotify®

LIVE DEMO

Data limitations & Bias

- The Medium class has a higher number of samples compared to the Low and High classes which could influence the model's behavior. This could lead to the model being more biased toward predicting the Medium class .
- Another limitation in the dataset is that each genre is represented by a fixed number of tracks—1000 per genre. While this helps to standardize the data across genres, it might not fully capture the diversity or richness of each genre.

Call to Action And Future work

Model Improvement: Enhance accuracy through tuning, additional feature engineering, and addressing dataset biases.

Exploring New Models: Investigate alternative machine learning models (e.g., neural networks).

Conclusions

This project showcases the use of machine learning to classify song danceability, predicting whether a song's danceability falls into low, medium, or high categories based on various audio features. Tableau visualizations enhance this analysis by exploring music trends and genre-specific patterns.

Thank You