STAT 766 Final Project Presentation

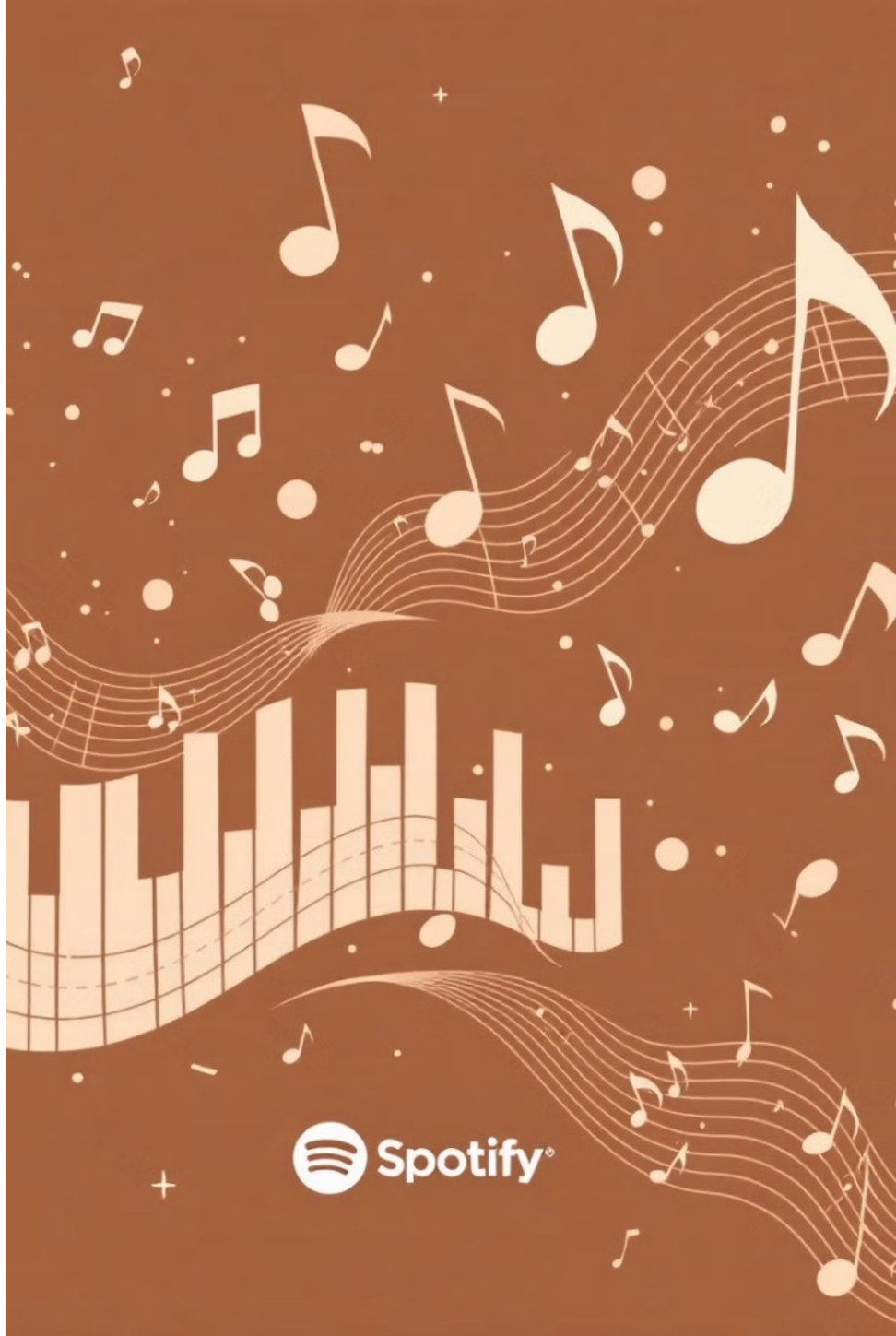
Spotify Music Classification

This project develops a robust machine learning pipeline to classify Spotify tracks into genres based on audio features. Using a comprehensive dataset of 30,000 songs, we explore data preprocessing, feature engineering, and model training to achieve accurate music genre classification.

Our study contributes to music information retrieval by advancing feature engineering techniques, benchmarking model performance, and developing a scalable pipeline for large-scale music classification tasks.

Student: Kai-Lin Wang

2024.12.05



Problem Statement

The problem we are addressing is the classification of music tracks based on their audio features. Specifically, we aim to predict categories such as genre based on various audio attributes like tempo, key, loudness, and danceability.

Spotify, with its vast library, provides a great opportunity to apply such classification models.

Significance of the Problem

1. Music Recommendation: Accurate classification helps improve music recommendation systems, making them more personalized and relevant.

2. **Music Discovery**: By categorizing tracks efficiently, we can introduce users to music they may not have encountered otherwise, enhancing their listening experience.

3. Industry Application: For the music industry, understanding trends and categorizing music effectively can help in market analysis, artist promotion, and playlist curation.

Data Preprocessing and EDA

1

Data Cleaning

Removed irrelevant features and extracted year from album release date

2

Feature Engineering

Created new features from existing data to improve model performance

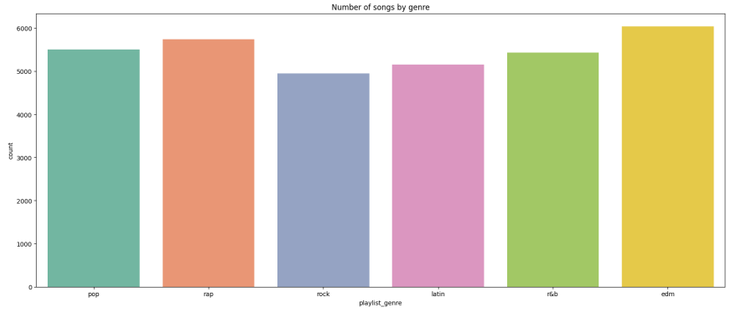
3

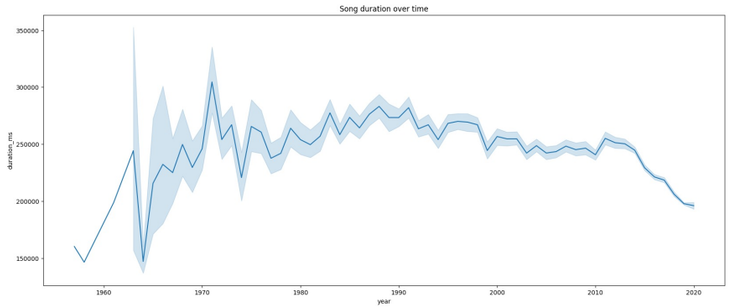
Exploratory Analysis

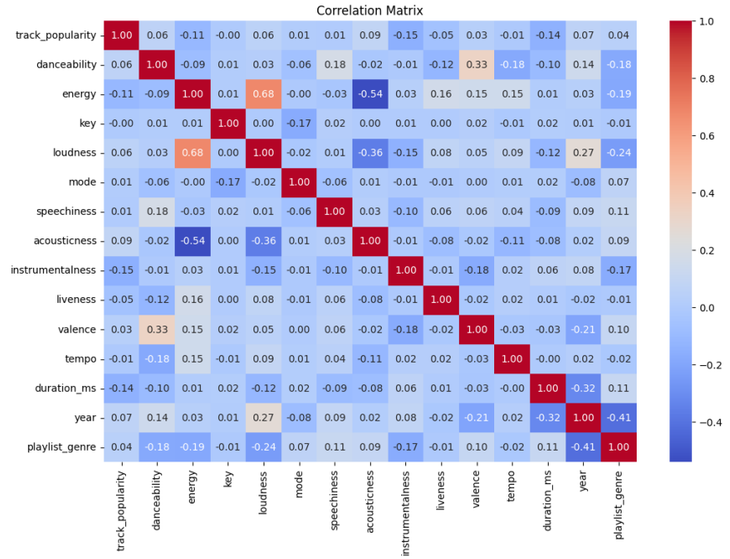
Generated visualizations to understand distribution of genres, popularity, and audio features

The preprocessing phase laid the foundation for our analysis, ensuring data quality and relevance. Our exploratory data analysis revealed insights into the dataset's characteristics, guiding our subsequent modeling approaches.

Key Dataset Insights







Genre Distribution

Visualization revealed the relative prevalence of different music genres in the dataset

Popularity Trends

Analysis of song popularity over time showed interesting patterns in listener preferences

Feature Correlations

Correlation matrix highlighted relationships between various audio features

Machine Learning Models

Random Forest

Ensemble learning method known for handling high-dimensional data and robustness to overfitting

XGBoost

Optimized gradient boosting algorithm for enhanced classification accuracy

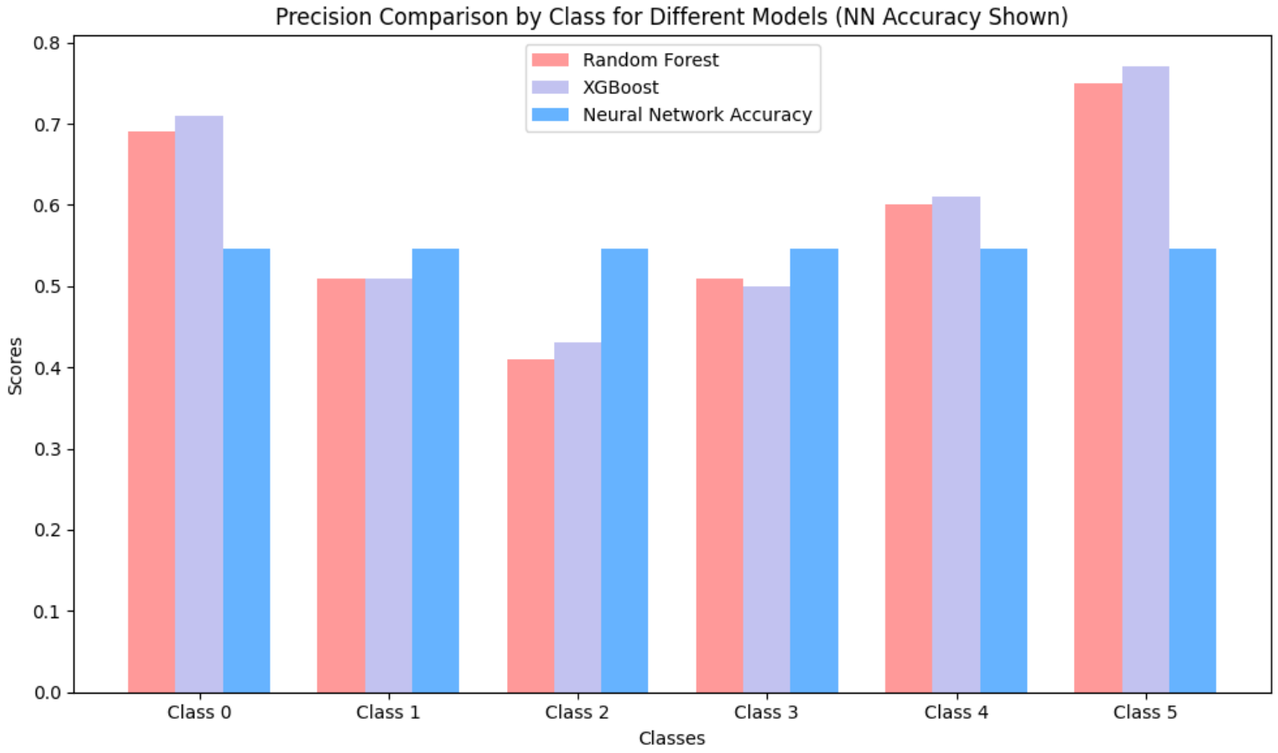
Neural Network

Deep learning approach to capture intricate patterns in the data

We implemented three distinct machine learning models to classify Spotify tracks into genres. Each model offers unique strengths in handling complex audio feature data.

- edm = 0
- latin = 1
- pop = 2
- r&b = 3
- rap = 4
- rock = 5

Model Performance Comparison



0.58

Random Forest

Accuracy achieved with optimized hyperparameters

0.591

XGBoost

Highest accuracy, outperforming other models

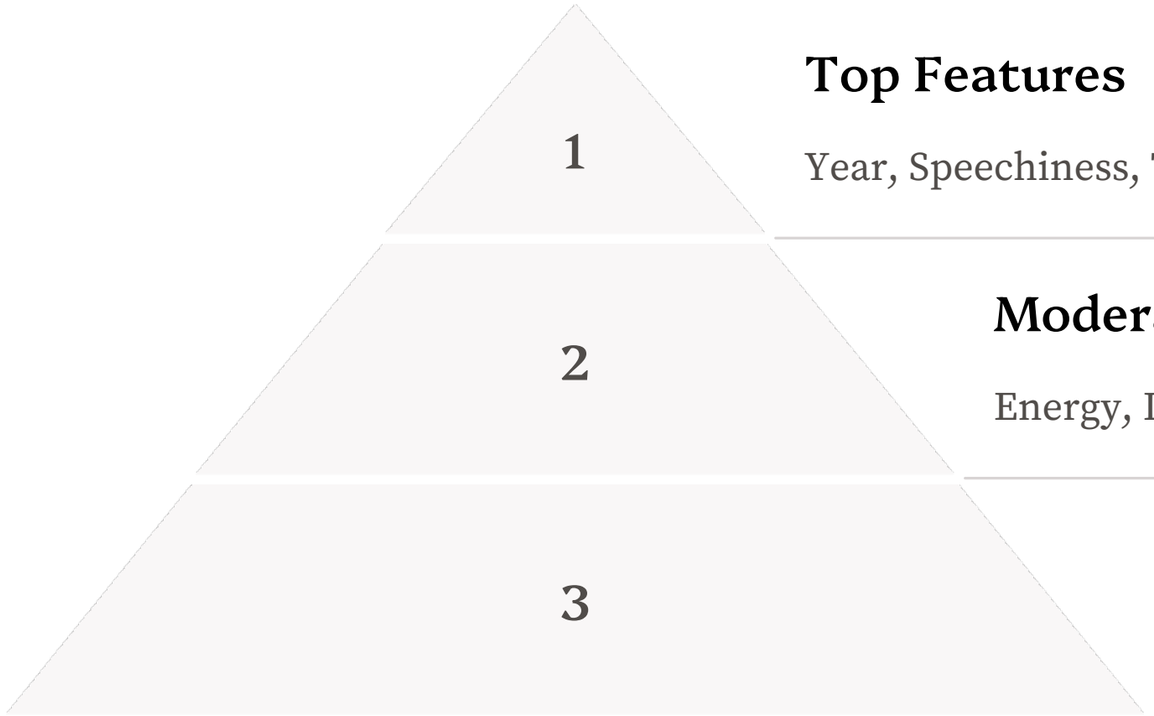
0.546

Neural Network

Test accuracy, showing potential for improvement

XGBoost demonstrated superior performance in genre classification, highlighting its effectiveness in handling complex, multi-class tasks. The Random Forest model also showed strong results, while the Neural Network approach indicates potential for further optimization.

Feature Importance Analysis



Top Features

Year, Speechiness, Tempo, Danceability

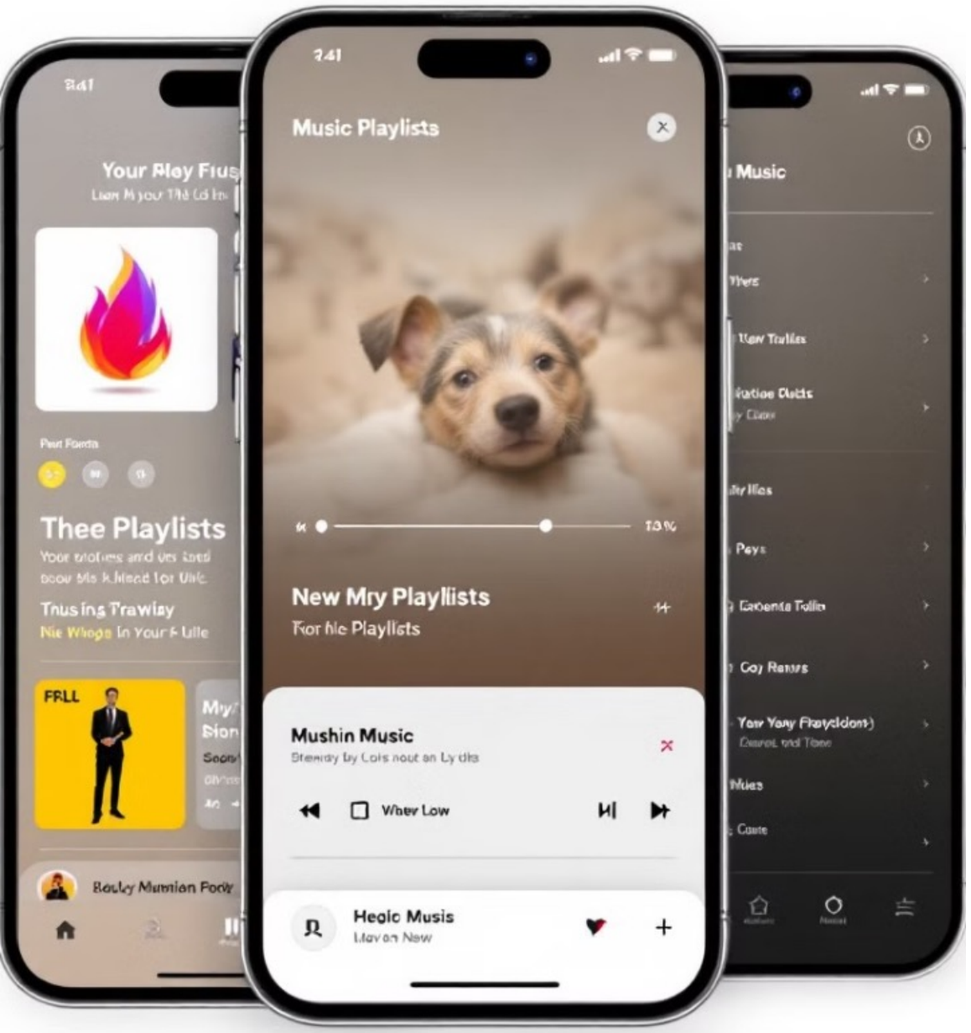
Moderate Impact

Energy, Loudness, Valence

Lower Influence

Mode, Key, Duration

Our analysis revealed that certain features played a crucial role in genre classification. Understanding feature importance helps refine models and provides insights into the key elements that define music genres.



Key Findings and Implications

XGBoost Superiority

Demonstrated best performance in genre classification

Feature Engineering Impact

Careful feature selection significantly improved model accuracy

Temporal Relevance

Release year proved crucial in genre prediction

Scalable Pipeline

Developed approach applicable to large-scale music datasets

Our findings have significant implications for music streaming platforms, enabling more accurate personalized recommendations, playlist curation, and music discovery features.

Future Directions

1

Advanced Deep Learning

Explore convolutional and recurrent neural networks for improved accuracy

2

Additional Features

Incorporate lyrics and contextual information for richer analysis

3

Cross-Platform Integration

Extend the model to classify music across different streaming platforms

4

Real-Time Classification

Develop systems for instant genre classification of new releases

The future of music classification holds exciting possibilities. By continuing to refine our approaches and incorporate new data sources, we can further enhance the accuracy and applicability of genre classification models.

STAT 766 Final Project Presentation

Spotify Music Classification

Thanks for your listening

Student: Kai-Lin Wang

2024.12.05