Music Streaming and Track Genre Prediction

Advanced Data Analysis Project

Victor Regly

University of Lausanne, HEC Lausanne

August 30, 2024

Outline

Introduction

- Music recommendation systems rely on accurate genre classification to enhance user experience.
- Challenges in genre classification include:
  - High diversity in musical styles.
  - Large volume of data with 114 genres.
- This project explores both supervised and unsupervised learning methods to predict music genres.

Research Question and Literature Review

Current Challenges:

- High dimensionality with 114 genres.
- Class imbalance across genres.
- Overlapping characteristics between genres.

Research Objectives:

- Develop a robust genre prediction model.
- Apply feature engineering and dimensionality reduction techniques.

Relevant Literature:

- Discusses various machine learning models for genre prediction.
- Importance of feature selection and handling class imbalance.

Dataset Description

- The dataset includes 114,000 entries from Spotify API.
- 21 columns with various features such as:
  - `danceability`, `energy`, `tempo`, `valence`.
  - `acousticness`, `instrumentalness`, `liveness`.
  - Metadata: `track_id`, `track_name`, `artists`, `album_name`.
- Target variable: `track_genre` categorizing each track into 114 distinct genres.

Exploratory Data Analysis (EDA)

- EDA was conducted to understand data distribution and relationships.
- Key findings:
  - Popularity skews towards lower scores.
  - Typical song duration between 3 to 6.5 minutes.

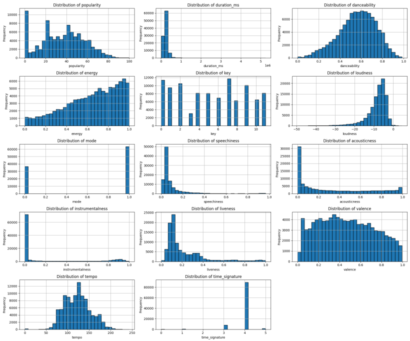

Figure: Distribution of Numerical Features

EDA: Popularity and Danceability by Genre

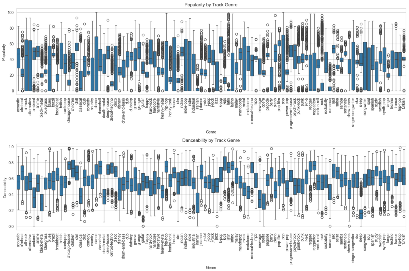

Figure: Popularity, Danceability by Genre

- Significant variability in popularity and danceability across genres.
- Genres like "pop" and "latin" are more popular and danceable.
- Highlights the complexity of genre classification.

Correlation Matrix of Features

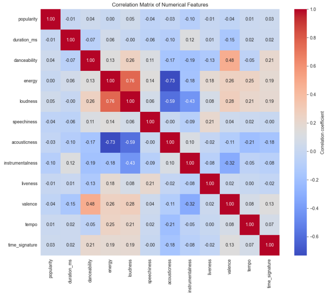

Figure: Correlation Matrix of Features

- Strong positive correlation between energy and loudness (0.76).
- Acousticness negatively correlated with energy (-0.73) and loudness (-0.59).
- Understanding these correlations is key to feature selection.

Methodology Overview

- Data preprocessing: Cleaning, normalization, encoding, and splitting.
- Supervised learning models: Logistic Regression, Random Forest, Neural Networks, XGBoost.
- Feature engineering and dimensionality reduction: RFE, PCA, t-SNE.
- Unsupervised learning: K-Means, Hierarchical Clustering, GMM.
- Ensemble methods: Stacking, Bagging, Boosting.

Data Preprocessing

- Handled missing values and normalized data using `StandardScaler()`.
- Categorical variables encoded with `LabelEncoder()`.
- Dataset split into 80% training and 20% testing sets to ensure robust model evaluation.

Supervised Learning: Model Performance

Table: Supervised Model Results

Model	Accuracy	Comments
Logistic Regression	0.2166	Baseline logistic regression with moderate accuracy.
Logistic Regression (L2 Regularization)	0.2169	L2 regularization slightly improved the model's performance.
Random Forest Classifier	0.3124	Significant improvement over Logistic Regression.
Random Forest (Grid Search)	0.3105	Grid search tuning didn't drastically improve performance.
XGBoost	0.3261	Better accuracy compared to Random Forest.
XGBoost (Tuned)	0.3291	Further tuning led to the best performance so far.
Neural Network	0.3131	Outperformed Random Forest but behind tuned XGBoost.
Neural Network (Tuned)	0.3328	Highest accuracy among all models.
Random Forest (Top 13 Features)	0.3133	Feature selection maintained decent accuracy with reduced features.
Random Forest (Top 16 Interaction Features)	0.2882	Including interaction features didn't improve accuracy.
Random Forest (PCA)	0.2047	PCA reduced accuracy significantly.

Feature Engineering and Dimensionality Reduction

- Applied Recursive Feature Elimination (RFE) to identify the most important features.
- Used Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction.
- These techniques improved model performance by reducing feature space and enhancing interpretability.

Hierarchical Clustering

- Hierarchical clustering grouped music genres based on shared characteristics.
- Successfully identified clusters that align with musical and cultural similarities.
- Limitations included highly imbalanced clusters and high computational cost.

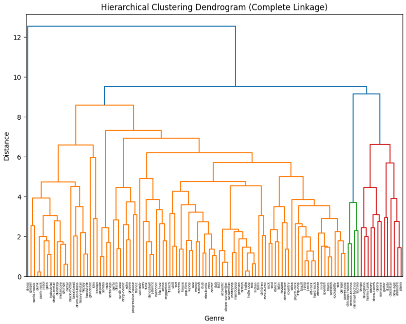

Figure: Hierarchical Clustering Results

Hierarchical Clustering: Neural Network Performance

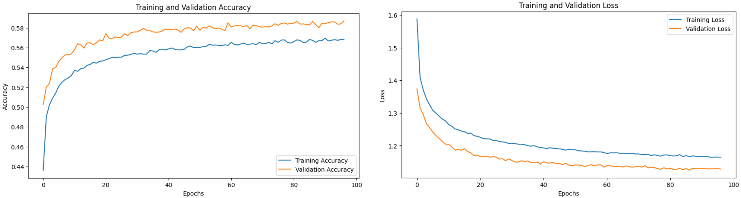

Figure: Neural Network Training and Validation Curves

- Training and validation curves show the model's learning and generalization.
- Close alignment of curves indicates effective generalization with minimal overfitting.

Hierarchical Clustering: Analysis and Issues

- Imbalanced clusters, some containing many genres, others very few.
- High computational cost and difficulty in determining the optimal number of clusters.
- Overlapping genres are not well-handled, leading to ambiguous cluster definitions.

K-Means Clustering

- Applied K-Means to optimize clustering based on Elbow Method and Silhouette Score.
- Found optimal clusters but had poor separation in actual genre prediction.
- Transitioned to GMM to handle genre overlaps and improve cluster flexibility.

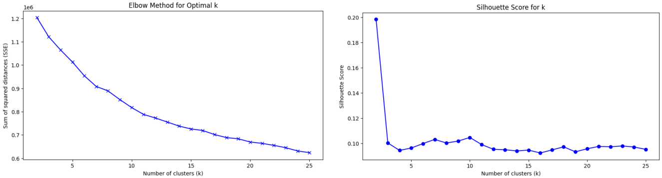

Figure: Elbow and Silhouette Graphs for K-Means

Gaussian Mixture Models (GMM)

- Transitioned to GMM due to limitations with K-Means.
- GMM provided soft cluster assignments, allowing for overlap and better handling of genre complexities.
- Improved the flexibility in modeling genre overlaps.

Stacked Ensemble Modeling

- Combined Neural Network, XGBoost, and Random Forest models.
- Meta-classifier (Logistic Regression) used to make final predictions.
- Achieved high accuracy of 99.26%, showing ensemble model's strength.

Ensemble Model Performance Analysis

- High overall accuracy of 99.26
- Consistent performance across most clusters, with high precision and recall.
- Slightly lower performance in some clusters suggests areas for further refinement.

Conclusion

- Predicting music genres is complex due to overlapping features and high data dimensionality.
- Ensemble methods and neural networks showed the best performance for supervised learning.
- Unsupervised learning highlighted the difficulty of genre classification.

Future Work

- Incorporate more nuanced audio features and metadata.
- Explore hybrid models combining both supervised and unsupervised techniques.
- Investigate deep learning models further, especially in handling large, high-dimensional data.

Thank You!

Questions?