Machine Learning Analysis of Musical Patterns: Classification and Regression on Spotify Dataset

TZU-CHIEH CHAO

Department of Engineering Education

Master of Science in Applied Data Science

University of Florida

Abstract—This project conducts a thorough analysis of musical features by utilizing machine learning techniques on the Spotify dataset. The study has two main analytical goals: classifying songs into high or low danceability categories and performing regression analysis to predict valence scores, which indicate a track's positivity or happiness. For the classification task, logistic regression and random forest classifier models were employed and compared. In the regression task, linear regression was applied both with and without regularization methods, along with decision tree regressors. The models' performance was evaluated using standard metrics such as accuracy, precision, recall, and F1-score for classification, and mean squared error, mean absolute error, and R-squared for regression. The results showed that the random forest classifier was the most effective for classifying danceability, achieving an accuracy of 0.755, while linear regression provided the best predictive performance for valence, with an R-squared value of 0.36. This research adds to the growing field of computational music analysis and offers methodological insights into audio feature extraction and prediction, potentially improving music recommendation systems and content categorization methods.

Keywords—machine learning, regression, classification, random forest, decision tree , regularization, music recommendation systems

I. Introduction

Music streaming platforms like Spotify have significantly changed how people discover and interact with music, leading to the creation of large datasets related to musical works and listening habits. This data includes acoustic features obtained from audio analysis algorithms, which quantitatively assess musical traits such as danceability, energy, and valence, among others. These characteristics offer unique opportunities for applying machine learning techniques to analyze and predict musical properties.

This study explores the use of various machine learning models to classify danceability, which measures how suitable a track is for dancing based on a combination of musical elements, and to predict valence, which indicates the emotional positivity of a track. High valence scores are linked to positive feelings (like happiness and euphoria), while low scores are associated with negative emotions (such as sadness and anger).

A thorough understanding and accurate prediction of these attributes have significant implications for music recommendation systems, playlist creation, and music information retrieval. For instance, classifying songs by their danceability could help in crafting playlists for specific activities, while predicting valence could improve mood-based music suggestions.

Previous studies have demonstrated the feasibility of using machine learning models to predict various musical attributes. However, there is a need for a systematic comparison of different models to determine which are most effective for specific prediction tasks. This research aims to address this gap by implementing and assessing multiple classification and regression algorithms.

In the classification task to predict whether a song has high or low danceability, logistic regression, and random forest classifiers were used. For the regression task focused on predicting valence scores, linear regression, regression with regularization, and decision tree regression were utilized.

Through a comparative analysis of these models' performance, this study aims to identify the most effective strategies for each task and clarify the challenges involved in predicting these musical attributes.

II. Methodology

A. Dataset Description

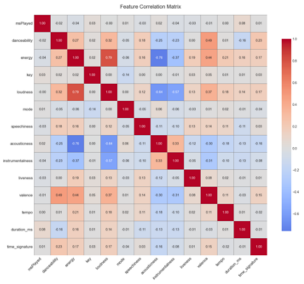
The Spotify dataset comes from Kaggle and includes music features and metadata for 10,080 songs. The features consist of danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, tempo, key, and valence, as well as metadata such as song title, artist, album, and genre. In the classification task, the target variable is danceability, while for the regression task, the target variable is valence.

B. Data Cleaning

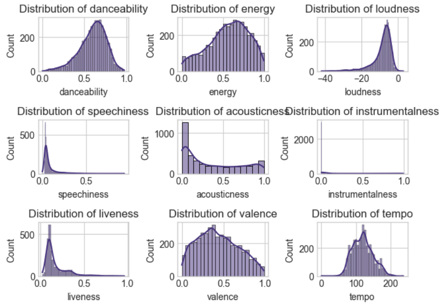
The "genre" column exhibits the highest incidence of missing values, with a total of 1,500 entries lacking data. Consequently, I have opted to remove the rows associated with these missing values, which concurrently addresses the issue of absent data in other features. Subsequently, I employ the unique "id" feature to identify and eliminate any duplicate entries based on the "id" column.

C. *Visualization Analysis*

- Correlation Matrix - The correlation matrix reveals key relationships between music features. Energy and loudness show strong positive correlation (0.79), while acousticness negatively correlates with both (-0.76, -0.64). Danceability and valence share moderate positive correlation (0.49). Most features display weak correlations with key, mode, tempo, and song duration, indicating their relative independence.



- Distribution of Features - The figure illustrates the distribution of nine significant musical characteristics derived from the Spotify dataset. Danceability is characterized by a normal distribution with a mean around 0.6, whereas energy is distributed more uniformly throughout its spectrum. Loudness demonstrates a pronounced skew towards higher values, particularly near 0 dB. The features of speechiness, instrumentalness, and liveness exhibit a strong positive skew, with the majority of tracks reflecting low values. Acousticness is represented by a bimodal distribution, displaying peaks at both very low and very high values. Valence, which indicates musical positivity, approximates a normal distribution centered at 0.5. Lastly, tempo is also normally distributed, with a mean around 120 BPM, although it includes some outliers at elevated tempos.



D. *Model Development*

D.1 Classification Task

Initially, the classification dataset is divided into training and testing subsets utilizing the train_test_split function from the scikit-learn library. In this process, 20% of the dataset is designated for testing, and a random seed value of 42 is employed to ensure the reproducibility of the results. Additionally, the shuffle parameter is configured to True, thereby ensuring a random allocation of samples between the training and testing sets.

Following the division of the dataset, the preprocessing pipeline for the classification task involves the standardization of six acoustic features: energy, loudness, speechiness, acousticness, instrumentalness, and valence. Employing scikit-learn's StandardScaler within a ColumnTransformer framework, each feature is normalized to have a mean of zero and a variance of one. This standardization is essential due to the varying scales of the features; for example, loudness, which is quantified in decibels, spans a range from -60 to 0, while the other features are normalized within the interval of 0 to 1. Standardization ensures that all features contribute equally to the model training process, thereby reducing the likelihood that features with larger scales will dominate others during the learning process. Additionally, this approach improves convergence rates

To forecast danceability, which is classified as either high (1) or low (0) with a threshold of 0.7, two classification models were utilized:

1. Logistic Regression: This model employs a linear classification framework and integrates L1 and L2 regularization techniques. The liblinear solver was employed to enhance the efficiency of the binary classification process.

2. Random Forest Classifier: This model is grounded in an ensemble learning methodology, permitting the modification of several parameters, such as the number of estimators, maximum depth, and the minimum number of samples required for a split.

D.2 Regression Task

A subset of the dataset is first created, comprising only the relevant acoustic features, specifically energy, danceability, acousticness, instrumentalness, loudness, and valence. This refined dataset is then partitioned into training (80%) and testing (20%) sets. Subsequently, a preprocessing pipeline is established for the regression task, which involves standardizing five acoustic features (energy, danceability, acousticness, loudness, and instrumentalness) utilizing StandardScaler within a ColumnTransformer framework. This normalization process guarantees that features with varying scales contribute equally to the models predicting valence.

For predicting valence (musical positiveness), multiple regression models were implemented:

3. Linear Regression: A baseline model to establish fundamental performance metrics

4. Linear Regression with regularization: Linear Regression with L1/L2 Regularization to prevent overfitting.

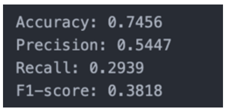
5. Decision Tree Regressor: A non-linear approach capable of capturing complex relationships between features

III. Evaluation

A. Classification

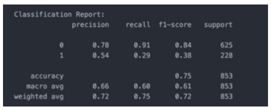
A.1 Model Performance

Logistic Regression Model : Achieved an accuracy of 74.56% in categorizing songs into high and low danceability classifications. The precision score of 0.5447 suggests a moderate level of reliability when the model predicts a song as highly danceable, while the recall score of 0.2939 indicates that the model successfully identifies only approximately 29% of all songs that are highly danceable. Consequently, the F1-score of 0.3818 highlights the disparity between precision and recall. Optimized through hyperparameter tuning, with the best performance achieved using L1 regularization (Lasso) and a relatively strong regularization strength (C=0.1).



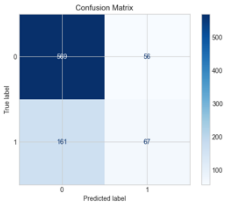
```
Accuracy: 0.7456
Precision: 0.5447
Recall: 0.2939
F1-score: 0.3818
```

The classification report offers a comprehensive analysis by class. For songs categorized as low danceability (class 0), the model demonstrates strong performance, with a precision of 0.78, a recall of 0.91, and an F1-score of 0.84 based on 625 samples. Conversely, the model's performance on high danceability songs (class 1) is significantly less robust, exhibiting a precision of 0.54, a recall of 0.29, and an F1-score of 0.38 across 228 samples.

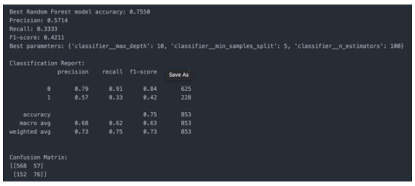


```
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.91      0.84       625
           1       0.54      0.29      0.38       228

    accuracy                           0.75       853
   macro avg       0.66      0.60      0.61       853
weighted avg       0.72      0.75      0.72       853
```

The confusion matrix further substantiates the issue of class imbalance, revealing 569 true negatives and 67 true positives, in contrast to 56 false positives and 161 false negatives. This data indicates the model's propensity to classify songs as having low danceability, demonstrating effective performance on the majority class while encountering difficulties with the minority class.



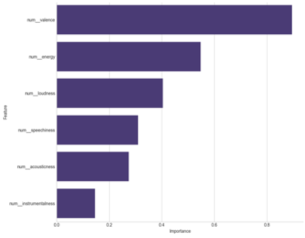
Random Forest Classifier: Achieved an accuracy of 75.50% in the classification task related to danceability, marginally surpassing the performance of the logistic regression model. It demonstrated a precision of 0.5714 and a recall of 0.3333, indicating an enhanced equilibrium in predictive performance, which culminated in a higher F1-score of 0.4211. The optimal hyperparameters identified for the model included a maximum tree depth of 10, a minimum samples split of 5, and 100 estimators. The classification report indicates a stronger performance for the majority class, characterized by low danceability, with a precision of 0.79 and a recall of 0.91. Conversely, the model's performance on high danceability songs remains challenging, although it has shown improvement relative to the logistic regression model. The confusion matrix illustrates 76 true positives and 568 true negatives, alongside 57 false positives and 152 false negatives.



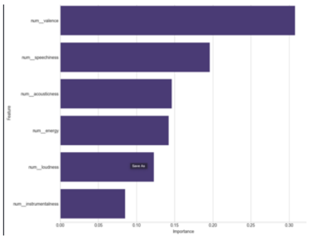
```
Best Random Forest model accuracy: 0.7550
Precision: 0.5714
Recall: 0.3333
F1-score: 0.4211
Best parameters: {'classifier__max_depth': 10, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100}

Classification Report:
              precision    recall  f1-score   Save As

           0       0.79      0.91      0.84       625
           1       0.57      0.33      0.42       228

    accuracy                           0.75       853
   macro avg       0.68      0.62      0.63       853
weighted avg       0.73      0.75      0.73       853


Confusion Matrix:
[[568  57]
 [152  76]]
```

A.2 Feature Importance

Logistic Regression Model: Shows valence as the dominant predictor of danceability, followed by energy and loudness. Speechiness and acousticness have moderate importance, while instrumentalness contributes least to the classification decision.



Random Forest Classifier : Also identifies valence as the most influential predictor of danceability, followed closely by speechiness. Acousticness and energy show moderate importance, while loudness and instrumentalness contribute least to the classification decision.

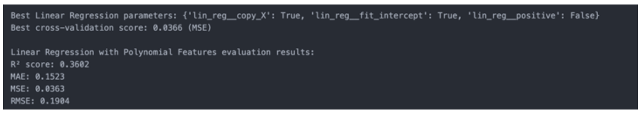


B. Regression

B.1 Model Performance

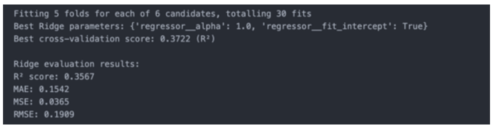
Linear Regression : The model demonstrated a moderate level of performance, as evidenced by an R^2 score of 0.3602, which indicates that it accounts for approximately 36% of the variance in valence values. Optimization of the model was achieved through the implementation of parameters such as fit_intercept=True and copy_X=True. The error metrics reveal a Mean Absolute Error (MAE) of 0.1523, a Mean Squared Error (MSE) of 0.0363, and a

Root Mean Squared Error (RMSE) of 0.1904, suggesting a reasonable yet limited capacity for predictive accuracy concerning this subjective musical attribute.



```
Best Linear Regression parameters: {'lin_reg__copy_X': True, 'lin_reg__fit_intercept': True, 'lin_reg__positive': False}
Best cross-validation score: 0.0366 (MSE)

Linear Regression with Polynomial Features evaluation results:
R² score: 0.3602
MAE: 0.1523
MSE: 0.0363
RMSE: 0.1904
```

Linear Regression (Ridge) : Achieved an R^2 score of 0.3567, explaining approximately 36% of the variance in musical positiveness. Performance metrics include MAE of 0.1542, MSE of 0.0365, and RMSE of 0.1909, showing slightly lower predictive accuracy than the base linear regression model despite employing regularization to prevent overfitting.



```
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Best Ridge parameters: {'regressor__alpha': 1.0, 'regressor__fit_intercept': True}
Best cross-validation score: 0.3722 (R²)

Ridge evaluation results:
R² score: 0.3567
MAE: 0.1542
MSE: 0.0365
RMSE: 0.1909
```

DecisionTreeRegressor(GridSearchCV): Underwent extensive hyperparameter tuning across 51,200 model configurations. The optimal model used squared error criterion with a maximum depth of 6, max features at 70%, minimum samples leaf of 4, and minimum samples split of 20. Performance metrics show an R^2 score of 0.2950, MAE of 0.1604, MSE of 0.0400, and RMSE of 0.1999, indicating lower predictive power than linear models for valence prediction.

B.2 Reflection on Low R-squared Values

1. Subjective Nature of Emotional Response: Musical emotion perception is highly individualized, influenced by cultural background, personal history, and aesthetic preferences, making objective prediction inherently difficult.

2.Non-linear Relationships: The association between acoustic properties and emotional responses is expected to exhibit significant non-linearity and is contingent upon contextual factors, which poses challenges even for decision tree models. Consequently, more sophisticated analytical methods should be employed to investigate the relationship between music and emotions.

IV. Conclusion & Business Insight

The machine learning models employed in this research exhibit differing levels of efficacy in forecasting musical characteristics. In the context of danceability classification, the Random Forest classifier achieved an accuracy of 75.50%, slightly surpassing the Logistic Regression model, which attained an accuracy of 74.56%. However, both models encountered challenges in accurately identifying songs with high danceability despite their overall commendable accuracy rates. Regarding valence prediction, all regression models demonstrated moderate performance, with Ridge Regression yielding the most favorable results, reflected by an R^2 value of 0.3567. Nonetheless, this indicates that the model accounts for only approximately one-third of the variance associated with emotional positivity.

These findings yield several valuable business insights for music streaming platforms:

1. Playlist Generation Strategy: The enhanced efficacy of ensemble methods in the classification of danceability indicates that playlist generators focused on activity should consider utilizing these more sophisticated models when assembling content for dance-oriented events.

2. Targeted Advertising Opportunities: Songs that are expected to have high danceability can be effectively matched with ads for dance clubs, fitness items, or energy beverages, leading to contextually appropriate marketing opportunities.

3. User Experience Personalization: Platforms might create flexible interfaces that adjust according to the anticipated emotional tone of the music being played—using brighter UI features for upbeat songs and more muted ones for downbeat tracks.

4. Recommendation System Enhancement: The limited predictive ability of acoustic features indicates that recommendation systems ought to integrate content-based elements with collaborative filtering to address the shortcomings of relying solely on acoustic analysis.

5. Emotional Journey Mapping: Although valence predictions may not be highly accurate, they could still be useful for visualization tools that illustrate the emotional "journey" of playlists or albums, assisting artists in creating more purposeful listening experiences.