Predicting Music Popularity: A Machine Learning Approach Using Spotify Data

Shuo Jiang[a]
Computer Science, Huazhong University of Science and Technology, Wuhan, China

Keywords: Machine Learning, Computer Science, Artificial Intelligence, Deep Learning.

Abstract: In today's world, with the continuous advancement and application of streaming technologies, music has become ubiquitous and is increasingly integrated into the daily lives of people. This paper examines the application of machine learning algorithms for predicting music popularity through an extensive dataset sourced from Spotify, comprising 114,000 songs recorded over two decades. Traditional methods of predicting song success have often been subjective and inaccurate; however, advancements in artificial intelligence (AI) offer new avenues for improvement. This paper employed three machine learning models‒Random Forest Regressor, Simple Linear Regression, and Gradient Boosting Machines‒to analyze various audio features and their influence on song popularity. The Random Forest Regressor surfaced as the most effective model, capturing complex relationships within the data and achieving a respectable R^2 score. The findings highlight key predictors of popularity, including danceability, energy, and loudness, while also revealing challenges in accurately forecasting songs at both ends of the popularity spectrum. This research highlights the significance of incorporating various elements, including marketing tactics and social media engagement, in addition to audio characteristics, to improve predictive accuracy. Ultimately, the study showcases the capability of machine learning methods in grasping the intricacies of music popularity dynamics.

1 INTRODUCTION

Music genres, much like trends, continually evolve in response to changing times and public tastes. The popularity of songs can fluctuate not only year by year but also month by month, making the prediction of music popularity a compelling area of study. Traditional predictive methods have often been subjective and data-deficient, leading to inaccuracies. However, advancements in Artificial Intelligence (AI) have introduced high-performance algorithms that enhance prediction accuracy and adaptability, making AI increasingly relevant in this field.

Recent progress in machine deep learning, especially with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has significantly improved AI capabilities, particularly with large-scale data processing. Reinforcement Learning (RL) has also shown remarkable potential in autonomous decision-making. The success of AlphaGo exemplifies RL's impact, contributing to the broader adoption of AI across diverse domains, including music.

For instance, omar et al. presents a categorization of AI techniques employed in algorithmic music composition., which focuses on the automatic generation of music by computer systems. The application of AI in this field involves utilizing various techniques as primary tools for creating compositions (Lopez-Rincon, 2018); In a heterogeneous network framework, Huan et al. organized the digraph set of track characteristics into multiple clusters, maximizing the diversity within each cluster, ensuring that the digraphs within each cluster are maximally isomorphic. By focusing similarity searches within the most relevant cluster to the target user, this approach enhances the efficiency of music recommendations by providing a sufficient selection of applicable tracks (Wang, 2022); Artistic style transfer is a fascinating application of generative AI that merges the content of one image with the artistic style of another, creating distinctive and imaginative visual artworks. Jonayet et al. present a

[a] https://orcid.org/0009-0006-7267-2908

Jiang, S.
Predicting Music Popularity: A Machine Learning Approach Using Spotify Data.
DOI: 10.5220/0013330000004558
In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 324-328
ISBN: 978-989-758-738-2

new approach to style transfer that employs CNNs (Miah, 2023); Gauri et al. summarized the research on using artificial intelligence technologies to filter, diagnose, monitor, and disseminate information about COVID-19 through human audio signals.

This overview will help develop automated systems to support COVID-19 related efforts to utilize non-invasive and user-friendly biosignals in human non-verbal and verbal audio (Deshpande, 2022); One of the most important directions for this is the prediction of music popularity, and here are some examples: HuaFeng et al. developed a model for predicting song popularity that combines multimodal feature fusion with LightGBM. The model consists of a LightGBM component, a multimodal feature extraction framework and a logistic regression component (Zeng, 2022); Notably, the research by Seon et al. empirically examined how acoustic features enhance the likelihood of songs reaching the top 10 on the Billboard Hot 100, analyzing data from 6,209 unique songs that appeared on the chart between 1998 and 2016, with a particular emphasis on acoustic features supplied by Spotify (Kim, 2021); In the research by Bang Dang et al., the paper focuses on predicting the rankings of popular songs for the next six months. The dataset, used for the Hit Song Prediction problem in the Zalo AI Challenge 2019, includes not only songs but also details like composers, artist names, release dates, and more. The paper advocates for treating hit song prediction as a ranking problem using Gradient Boosting techniques, rather than the typical regression or classification methods employed in previous studies. The optimal model demonstrated strong performance in predicting whether a song would become a top Ten dance hit versus lower-ranked positions (Pham, 2020).

Thanks to the robust development in this field, this paper also aims to employ AI algorithms for popularity prediction. To achieve this objective, the study utilizes extensive streaming data, including official metrics such as Spotify's track play counts and datasets from Kaggle relevant to the model. Experimental results validate the effectiveness of the proposed methods.

2 METHODS

2.1 Dataset Preparation

The Dataset which this paper picked was a Spotify Songs dataset that recorded 114,000 songs with their popularity, artists, genre, duration, etc.

These features can be used to predict a song's popularity and also to explore how these features influence that popularity. Additionally, this study conducted an online search for streaming play counts and popularity data for singles from 2004 to 2024. To account for regional differences, data was collected primarily from Spotify, YouTube Music, and QQ Music. These datasets were used as another critical source of information. Utilizing these datasets, the study conducts a regression task to examine the relationship between play counts and a song's popularity.

After cleaning the data, this paper selected features that were not popularity to become the independent variables. Then, were selected only the popularity as out dependent variable since its the target that this study aims to predict. In terms of data preprocessing, this paper conducted normalization training. To properly evaluate the model's performance, it's important to split the dataset into training and testing sets. This paper makes use of the "train-test-split" function from the “sklearn.model_selection” module, allocating 80% of the data to the training set and 20% to the testing set.

2.2 Machine Learning Models-Based Prediction

About the models this study chosen, this paper selected three different models. They are Random Forest Regressor(RF),Gradient Boosting Machines (GBM) and Simple Linear Regression.

2.2.1 Random Forest

Firstly, RF shown in Figure 1 is an ensemble method that constructs multiple decision trees and merges their results. It leverages bootstrapping and feature randomness to enhance model performance and reduce overfitting. It's Methodology including Ensemble Construction which generates multiple decision trees using bootstrap samples from the training data. Besides, each tree is trained on a unique subset of the data, which aids in minimizing variance and preventing overfitting. The reasons of why this study chose it are as follows: 1. powerful ensemble learning method; 2. It is capable of effectively handling both linear and non-linear relationships; 3. it offers robustness against overfitting, especially in datasets with many features.

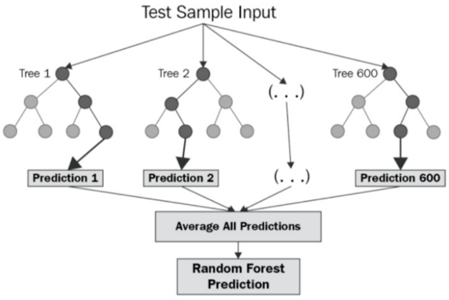

Figure 1: The structure of the RF (Muchisha, 2021).

2.2.2 Linear Regression

Simple Linear regression is important in modeling, encompassing model specification, model estimation, statistical inference, model diagnostics and prediction (Su, 2012). It is a commonly employed method in statistical modeling and machine learning for forecasting a continuous response variable using one or more input variables. Its objective is to uncover the linear relationship between the variables by minimizing the discrepancy between observed and predicted values. The model could be showed as:

$$Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta p Xp + \epsilon. \quad (1)$$

$\beta 0$ is the intercept, β_i are the coefficients, x_i are independent variables, Yis the dependent variable and ϵ is the error term. Its target is to estimate β that best fits the observed data. A widely used approach for estimating coefficients is Ordinary Least Squares (OLS). Mathematically, the objective is to minimize:

$$\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \quad (2)$$

The reasons this study chose Simple Linear Regression was because of its simplicity and interpretability.

2.2.3 Gradient Boosting Machines

Lastly, this paper selected Gradient Boosting Machines (GBM). GBM encompass a group of effective machine-learning methods that have achieved significant success across various real-world applications. They can be tailored to meet specific application needs, including the ability to be trained with different loss functions (Natekin, 2013). It is a robust ensemble learning technique that constructs predictive models by sequentially incorporating weak learners and enhancing their performance using gradient descent. The GBM algorithm follows a boosting framework where the model is constructed in a sequential manner. In every iteration, a new weak learner is fitted to the residuals from the current ensemble's predictions. The final model aggregates all weak learners, with each weighted according to its effectiveness. Mathematically, the model can be showed as:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \quad (3)$$

$F(x)$ is the final model prediction, $h_m(x)$ represents the $m-th$ weak learner, and α_m is the weight associated with the $m-th$ learner.GBM employs gradient descent to minimize a specified loss function. In each boosting iteration, it evaluates the gradient of the loss function relative to the current model's predictions, subsequently training a new weak learner to approximate this gradient, thereby effectively diminishing residual errors.

It's update rule for the model can be showed as:

$$F_{m+1}(x) = F_m(x) + \eta \cdot F_{m+1}(x) \quad (4)$$

η is the rate of learning, which regulates the contribution of each new learner. Of course, to prevent overfitting and enhance generalization, GBM incorporates regularization technique such as pruning the decision trees and incorporating constraints on tree depth or leaf nodes.

GBM offers several advantages: it builds models sequentially and its high predictive accuracy. Also, GBM offers insights into feature importance, which helps in understanding and interpreting the model.

3 RESULTS AND DISCUSSION

Regarding model performance, the final Random Forest Regressor was assessed on the test set, yielding promising results.

3.1 Random Forest Performance of Regression

The key performance metrics are as follows:

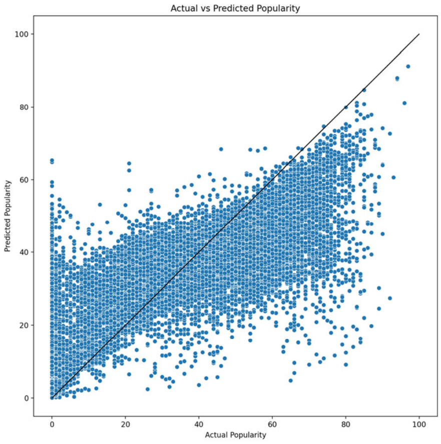

Figure 2: The prediction performance of the RF (Photo/Picture credit: Original).

From the Figure 2, the model achieved an R ^{2} score of 0.6137437848169063. In addition, an MSE of 190.61005900977344 was obtained.

High Popularity Predictions: The model exhibited a tendency to underpredict several popular songs. This discrepancy may arise from the unique traits or marketing strategies associated with these tracks, which were not fully captured by the audio features used in the model. As a result, the model may have overlooked important factors influencing a song's popularity, such as cultural context or promotional efforts.

Low Popularity Predictions: On the other hand, the model overpredicted the popularity of certain low-scoring tracks. This issue could be attributed to data noise or the misalignment of niche genres with mainstream metrics. Tracks from less popular genres may not align well with the features that typically drive popularity in more mainstream contexts, leading to inaccurate predictions. This highlights the need for a more nuanced approach to feature selection and model tuning, especially when dealing with diverse musical styles.

3.2 Learning Curve

Learning Curve is shown in Figure 3 follows:

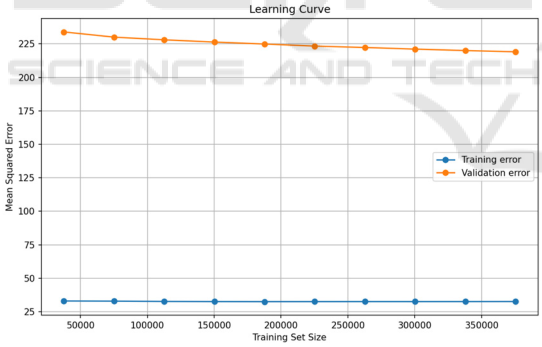

Figure 3: The learning curve (Photo/Picture credit: Original).

The training error is close to zero, but the validation error stays high. The validation error does not change much as the size of the training set increases.

3.3 Random Forest Performance of Classification

This study initially evaluated the model's performance with varying numbers of classes.

Figure 4 is the confusion matrix when the whole popularity is classified into 4 parts.

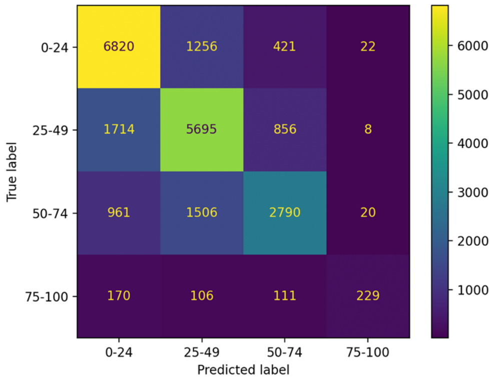

Figure 4: The confusion matrix of 4 parts (Photo/Picture credit: Original).

Figure 5 is the accuracy when the whole popularity classified into 3 parts.

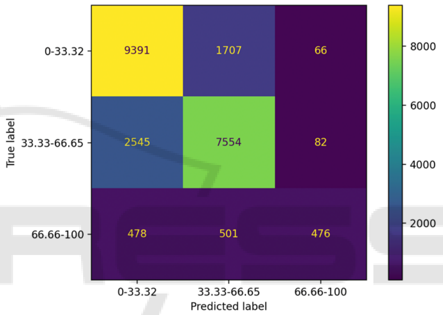

Figure 5: The confusion matrix of 3 parts (Photo/Picture credit: Original).

Figure 6 the confusion matrix when the whole popularity is classified into 2 parts.

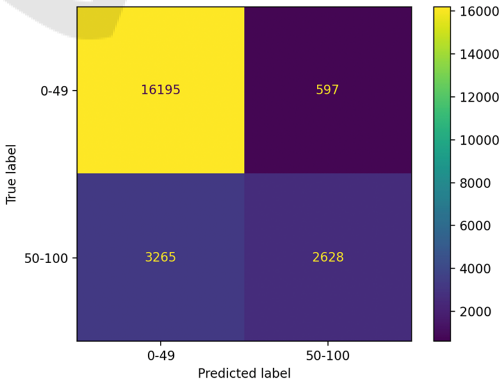

Figure 6: The confusion matrix of 2 parts (Photo/Picture credit: Original).

For feature importance and performance analysis, the results can be found in Table 1 and Figure 7.

Table 1: The performance of the model.

	Precision	Recall	F1-score	Support
1	0.76	0.84	0.80	11164
2	0.77	0.74	0.76	10181
3	0.76	0.33	0.46	1455
accuracy			0.76	22800
Macro avg	0.76	0.64	0.67	22800
Weighted avg	0.76	0.76	0.76	22800

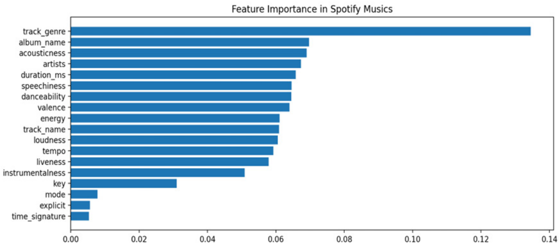

Figure 7: The feature importance (Photo/Picture credit: Original).

To sum up, while the model performed well overall, the error analysis revealed some challenges in accurately predicting songs at the extremes of the popularity spectrum. This indicates that while audio features are important, other factors like marketing efforts, artist reputation, lyrics, and social media presence may also impact popularity.

Additionally, this paper explored the use of a Random Forest Classifier for this task, which provides approximate popularity levels for each song and achieved acceptable accuracy. The analysis emphasized important features like energy, loudness, danceability and valence as crucial indicators of a song's success.

4 CONCLUSIONS

This study effectively built machine learning models to predict song popularity using data sourced from Spotify. After thoroughly evaluating several candidate models, this study ultimately preferred the Random Forest Regressor for its outstanding ability to capture the complex relationships between audio features and song popularity effectively. Its performance in modeling these complexities stood out, making it the ideal choice for the analysis. This model demonstrated strong performance in accuracy and achieved a notable R^2 score, reflecting its ability to account for a considerable amount of the variance in song popularity.

The analysis revealed that the model effectively identified patterns within the data, allowing for meaningful predictions. However, it also highlighted the challenges of accurately forecasting popularity for songs at both ends of the popularity spectrum. While audio features play a crucial role, the model suggests that other factors—such as marketing strategies, artist reputation and social media presence – may also significantly influence a song's success.

Overall, the research findings provide valuable insights into the dynamics of music popularity and underscore the potential of machine learning techniques in this field.

REFERENCES

Deshpande, G., Batliner, A., & Schuller, B. W. 2022. AI-Based human audio processing for COVID-19: A comprehensive overview. Pattern recognition, 122, 108289.

Kim, S. T., & Oh, J. H. 2021. Music intelligence: Granular data and prediction of top ten hit songs. Decision Support Systems, 145, 113535.

Lopez-Rincon, O., Starostenko, O., & Ayala-San Martín, G. 2018. Algorithmic music composition based on artificial intelligence: A survey. In 2018 International Conference on Electronics, Communications and Computers (CONIELECOMP) (pp. 187-193). IEEE.

Miah, J., Cao, D. M., Sayed, M. A., & Haque, M. S. 2023. Generative AI Model for Artistic Style Transfer Using Convolutional Neural Networks. arXiv preprint arXiv:2310.18237.

Muchisha, N. D., Tamara, N., Andriansyah, A., & Soleh, A. M. 2021. Nowcasting Indonesia's GDP growth using machine learning algorithms. Indonesian Journal of Statistics and Its Applications, 5(2), 355-368.

Natekin, A., & Knoll, A. 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

Pham, B. D., Tran, M. T., & Pham, H. L. 2020. Hit song prediction based on gradient boosting decision tree. In 2020 7th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 356-361). IEEE.

Su, X., Yan, X., & Tsai, C. L. 2012. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.

Wang, H. 2022. AI-Based Music Recommendation Algorithm under Heterogeneous Network Platform. Mobile Information Systems, 2022(1), 7267012.

Zeng, H., Yuan, Q., Guo, L., & Xu, S. 2022. Song popularity prediction model based on multi-modal feature fusion and LightGBM. In Proceedings of the 8th International Conference on Communication and Information Processing (pp. 28-32).