Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Optimal Machine Learning Algorithms for Predicting the Popularity of Songs

Zhuoyang Tao *

Hwa Chong International School, Bukit Timah, Singapore

* Corresponding Author Email: tao.zhuoyang@stu.hcis.edu.sg

Abstract. Predicting song popularity has become a hot topic of research in recent years due to its necessity. This study investigates the effectiveness of different machine learning models in predicting the popularity of songs on Spotify using audio features. By comparing Linear Regression, Random Forest, and K-Nearest Neighbors (KNN), the research aims to identify the most suitable algorithm for this task. A dataset containing key musical attributes such as danceability, loudness, energy, tempo, and valence was preprocessed through standardization and one-hot encoding. Model performance was evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). The results show that Random Forest outperforms the other models with the lowest prediction error and highest explanatory power. Additionally, feature importance analysis revealed that duration, speechiness, and emotional characteristics like energy and valence are more decisive in determining a song's popularity, whereas musical key and mode are less influential. The study concludes that while audio features offer valuable insights, external factors such as playlist placement and social media trends should be considered in future work to improve prediction accuracy.

Keywords: Spotify; song popularity; machine learning.

1. Introduction

Music has been one of the most common forms of entertainment. For listeners, they can enjoy their favorite melody to relieve stress, while for music creators, they rely on people buying their albums and songs in order to make a living. Today, music is a well-established and globally significant industry.

And all this popularity of music attributes to media streaming platforms to a large extent, for example, Spotify, YouTube Music, NetEase Cloud Music etc. They are making music more accessible for the public, just one touch away from their mobile devices. Among them, Spotify is now of the most widely used platform, having a number of 675 million monthly active users, with currently over 100 million songs available for streaming [1]. Spotify has a ranking of most popular songs lively updated, showing the hottest song in the world. The ranking to some extent shows the taste and preference of worldwide listeners. Hence, for music listeners around the world, they would like to have more music created that resembles songs in the ranking playlist, and for music creators, they would make more profit if they create music that meet the public's needs. Therefore, there shows a huge demand for an accurate prediction of possible popular hit music type, to facilitate personalized music recommendations, guide music production, and support decision-making in the music industry.

Predicting song popularity has become a hot topic of research in recent years due to its necessity. In order to predict song success effectively, academics have investigated a variety of machine learning techniques, such as ensemble learning, decision trees, deep learning, and linear regression, in light of the growing importance of data-driven decision-making in the music industry. For example, linear regression has been widely used as a baseline model in music analytics [2]. While it provides interpretability and a straightforward analysis of relationships between features, studies show that it often fails to capture non-linear patterns within data, leading to suboptimal predictions [3]. Random forest and gradient boosting techniques, such as XGBoost and LightGBM, have shown promising results in handling non-linear relationships in song popularity prediction. These models leverage feature interactions and provide better generalization compared to simple regression models. K-Nearest-Neighbors (KNN) is a non-parametric technique that uses music similarity to predict popularity. While useful for recommendation systems [4], when compared to more complex models like random forests or neural networks, KNN frequently performs worse and has trouble with large dimensional feature spaces. Neural networks like Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) architectures are examples of recent developments in deep learning that have shown promise in forecasting human involvement with music [5]. However, their need for large datasets and computational power limits their practicality in smaller scale analyses. In general, prior works are done focusing on a single machine learning method, but never focused on a comprehensive comparison between multiple algorithms, without further intention to find the optimal algorithm in terms of music popularity prediction.

This research aims to compare multiple models side by side, including: Traditional Regression (Linear Regression), Tree-Based Models (Random Forest) and Similarity-Based Models(K-Nearest-Neighbors) in Spotify song popularity prediction. This direct comparison provides a comprehensive understanding of which algorithm performs best for prediction.

2. Method

2.1. Dataset Preparation

In this study, the songs dataset provided by a dataset on Kaggle was used for analysis [6]. 900 songs were included in the original data; danceability, energy, key, loudness, mode, speechiness, acousticness, liveness, valence, tempo, and duration were among the numerical characteristics. The playlist genre categorization function was also added. The target variable for prediction is track popularity, a numerical score representing a song’s success on the platform, hence it should be a regression problem.

The preprocessing of the collected dataset consists of 3 parts. First, this study built up a pipeline. In order to eliminate bias toward big numerical values, like duration, all variables were brought to a comparable scale by standardizing numerical characteristics using StandardScaler. The categorical feature, playlist genre, was converted into numerical format using One-Hot Encoding. The dataset was then split into 90% training data and 10% test data to allow for sufficient model training while maintaining a separate evaluation set.

2.2. Machine-Learning Based Prediction Models

This study used 3 algorithms implemented by sklearn: Linear Regression, Random Forest, and K-Nearest-Neighbors (KNN). The evaluation metrics used are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-Squared (R^2) value.

2.2.1. Linear regression

A basic statistical technique for simulating the relationship between one or more independent variables (audio attributes) and a dependent variable (track popularity) is linear regression [7, 8]. It looks for the best-fitting line that reduces the discrepancy between actual and anticipated values, assuming a linear connection between variables. Even if linear regression offers a model that can be understood, it might not be able to fully represent intricate, non-linear relationships in the data.

2.2.2. Random forest

In order to increase predicted accuracy, Random Forest, an ensemble learning technique, builds several decision trees during training and combines their results [9, 10]. It improves resilience and lessens overfitting by averaging the predictions from many decision trees. In contrast to linear regression, Random Forest is especially good at managing high-dimensional data and may capture non-linear relationships between features. It is also feasible to determine which input variables have the greatest influence on the model's predictions by using random forest's ability to determine the relevance of each feature, which allows one to evaluate the contribution of each feature to the model's predictive accuracy.

2.2.3. K-Nearest-neighbors

K-Nearest neighbors is a distance-based algorithm that predicts the target variable based on the similarity between observations. It works by identifying the k most similar instances (neighbors) in the training set and averaging their values to make a prediction. KNN is simple and intuitive but may suffer from performance issues in high-dimensional spaces where distance calculations become less meaningful.

3. Results and Discussion

3.1. The performance of models

The results of this study shown in Table 1 provide an in-depth comparison of three machine learning models including Linear Regression, Random Forest, and KNN in predicting the popularity of songs on Spotify. The evaluation metrics used include MSE, RMSE, MAE, and R^2. These metrics allow for an objective assessment of each model's predictive accuracy and reliability.

Table 1. Performance metrics of different machine learning models.

Model	MSE	RMSE	MAE
Linear Regression	0.0301	0.1736	0.1432
Random Forest	0.0279	0.1670	0.1377
KNN	0.0390	0.1974	0.1532

According to the results, Random Forest performs better in terms of prediction than both KNN and Linear Regression. The Random Forest model exhibits the lowest error rates among the three models, with an MSE of 0.0279 and an RMSE of 0.1670, suggesting that it offers the most precise predictions of song popularity. Even if the model accounts for some of the data's volatility, its R^2 score of 0.1775 indicates that total explanatory power can still be increased.

Linear Regression, serving as a baseline model, performs moderately well with an MSE of 0.0301, an RMSE of 0.1736, and an MAE of 0.1432. Its R^2 score of 0.1117 suggests that the model is limited in its ability to explain the variability in song popularity, likely due to the linear nature of the model, which may not fully capture the complex relationships between audio features and popularity.

With an RMSE of 0.1974 and an MSE of 0.0390, KNN had the largest prediction error and fared the worst out of the three models. Furthermore, the model performs worse than a simple mean prediction, as evidenced by its negative R^2 score of 0.1496, which shows that it is unable to decipher any variance in the data. This outcome is probably the result of the dataset's high dimensionality, which hinders KNN's capacity to produce precise predictions because distance-based learning techniques are weakened by the curse of dimensionality.

3.2. Feature importance

The results shown in Fig. 1 show that the duration, speechiness, valence, energy and tempo of a song play relatively important roles in deciding the popularity of the song, while those like 'key' and 'mode' are not as decisive.

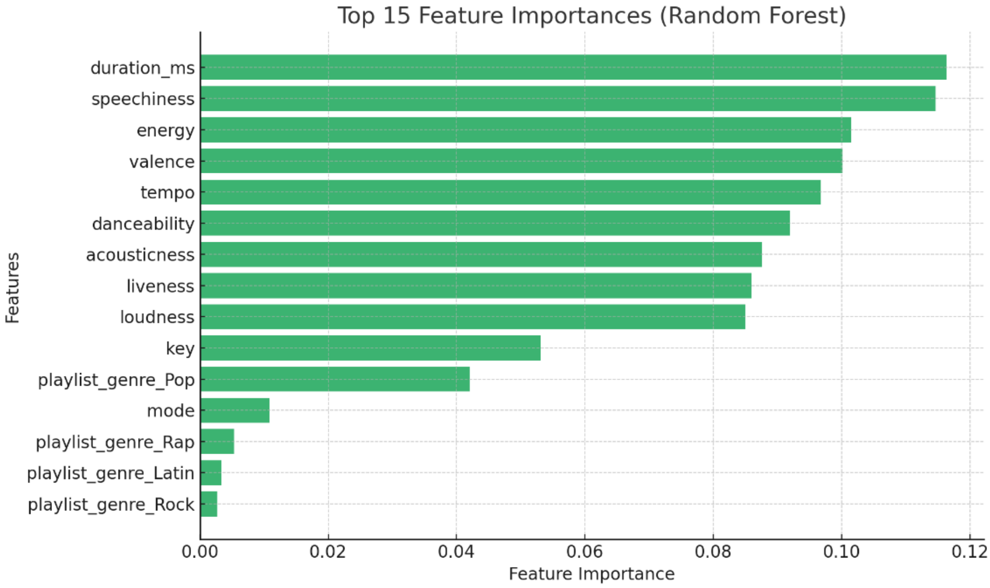

Fig. 1 Feature importance of each audio feature based on Importance Score (Picture credit : Original).

Duration is a measure of length of songs, long or short songs might follow different commercial patterns, e.g., radio-friendliness and TikTok virality. Duration's prominence suggests that song length correlates with audience engagement. Speechiness measures the presence of spoken words. Higher speechiness may align with rap/spoken genres, which have different engagement patterns. Valence represents musical positivity. Songs with higher valence may be perceived as happier tracks. Energy captures intensity and activity. High-energy songs might trend better in workout or party playlists. Tempo refers to bpm (beat per minute), and it can influence how danceable or engaging a song feels.

This feature importance check indicates that not all audio features contribute equally to popularity prediction. Duration, speechiness, and emotional characteristics (valence, energy) are key determinants. Features like "key" and "mode", which represent pitch structure and tonality, were less important. It provides the insight that popular songs on Spotify tend to share structural and emotional traits. However, only audio features were included—external drivers like playlist inclusion, artist popularity, or social media virality are not represented.

4. Conclusion

This study identified the most effective machine learning model for predicting Spotify song popularity using audio features. The results indicate that Random Forest outperforms Linear Regression and KNN, although all models demonstrate relatively low predictive accuracy due to various limitations. While Random Forest was the best performing model, its low R^2 score suggests that additional factors beyond audio features significantly influence song popularity. Feature importance check conducted in Random Forest showed that not all audio features contribute equally, duration, speechiness and emotional characteristics play more decisive roles than others, showing that popular songs share similar traits.

Future research should explore incorporating external variables such as playlist inclusions, social media engagement, artist popularity, and listener demographics to improve model accuracy. Additionally, experimenting with advanced machine learning techniques like Gradient Boosting, XGBoost, or deep learning architectures may enhance predictive performance. A larger and more diverse dataset could also contribute to better generalization across different song genres and musical trends.

References

[1] Spotify. About Spotify. Spotify, 2025. https://newsroom.spotify.com/company-info/

[2] Herremans D, Martens D, Sörensen K. Dance hit song prediction. Journal of New Music Research, 2014, 43(3): 291–302.

[3] James G, Witten D, Hastie T, Tibshirani R, Taylor J. Linear regression. In: An Introduction to Statistical Learning: With Applications in Python. Cham: Springer International Publishing, 2023: 69–134.

[4] Schedl M, Zamani H, Chen C W, Deldjoo Y, Elahi M. Current challenges and visions in music recommender systems research. International Journal of Multimedia Information Retrieval, 2018, 7(2): 95–116.

[5] Choi K, Fazekas G, Sandler M. A comparison of audio signal preprocessing methods for deep neural networks on music tagging. arXiv preprint, 2017. https://arxiv.org/abs/1709.01922

[6] RishabhPancholi1302. Spotify most popular songs dataset. Kaggle, 2024. https://www.kaggle.com/datasets/rishabhpancholi1302/spotify-most-popular-songs-dataset

[7] Anderson T W, Brown J R, Hall J W, Shephard R J. The limitations of linear regressions for the prediction of vital capacity and forced expiratory volume. Respiration, 1968, 25(2): 140–158.

[8] Nie L, Chu H, Liu C, Cole S R, Vexler A, Schisterman E F. Linear regression with an independent variable subject to a detection limit. Epidemiology, 2010, 21(4): S17–S24.

[9] Rigatti S J. Random forest. Journal of Insurance Medicine, 2017, 47(1): 31–39.

[10] Biau G, Scornet E. A random forest guided tour. Test, 2016, 25(2): 197–227.