International Journal of Basic and Applied Sciences, 14 (2) (2025) 711-720

SPC

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/gjywez76
Research paper

Machine Learning Approaches for Predicting Song Popularity: A Case Study in Music Analytics

Uppuluri Lakshmi Soundharya [1] *, Sasidhar Reddy Gaddam [2], Gogineni Krishna Chaitanya [3], T. L. Deepika Roy [3], Uppuluri Naga Lakshmi Madhuri [4]

[1] Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[2] Staff IT Software Engineer, Palo Alto Networks, Huntersville, North Carolina, USA
[3] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[4] Department of Computer Science and Engineering NRI Institute of Technology Pothavarapadu, Andhra Pradesh, India
**Corresponding author E-mail: mail2usoundharya@gmail.com*

Received: May 19, 2025, Accepted: June 21, 2025, Published: July 10, 2025

Abstract

Comprehending the aspects that impact song popularity has become crucial in the ever-changing music industry. This study explores the field of music popularity predictive modeling using the cutting-edge algorithms XGBoost and LightGBM. Predictive models were developed by the study using a large dataset that includes a variety of musical variables, such as song duration, tempo, lyrical content, and release year. To improve the models' predictive capacity, the study approach includes extensive work. To provide a thorough assessment of the algorithms' performance, the dataset is divided into training and testing sets. Additionally, the effectiveness of XGBoost and LightGBM forecasting music popularity is evaluated by a comparison analysis. To increase the prediction models' accuracy, hyperparameter optimization methods—specifically, Optuna—are used to fine-tune them. In addition, the study looks at feature importance, illuminating elements of music that, in the eyes of each algorithm, greatly add to its appeal. Using a rigorous cross-validation approach, the models are validated, and their generalization capabilities are shown. The performance metrics, which provide a comprehensive picture of the models' predicted accuracy, include mean absolute error, mean squared error, median absolute error, and R-squared. By providing a comparative analysis of two well-known machine learning methods for forecasting music popularity, this paper advances the rapidly developing field of music analytics. The results offer significant perspectives for professionals in the field and data scientists who are looking for efficient approaches to forecast music popularity across various genres.

Keywords: *Music Popularity Prediction; Machine Learning; XGBoost; LightGBM; Predictive Modeling; Feature Engineering; Hyperparameter Optimization; Data Analytics; Comparative Analysis; Song Characteristics; Genre Classification; Ensemble Models; Cross-Validation; Optuna; Data Preprocessing; Feature Importance; Regression; Music Analytics.*

1. Introduction

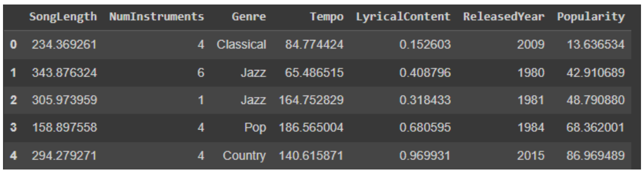
The modern music industry is a perfect area to apply state-of-the-art machine learning techniques since it possesses an unmatched amount of data. In this era of digital consumption, understanding the factors that contribute to a song's success is not only fascinating but also crucial for musicians, producers, and other industry participants. XGBoost and LightGBM. As our interactions with musical content continue to alter due to the digitization of music, it is effective and important.[21] The intricate relationships between various musical parts are often missed by traditional methods of assessing a song's potential. This work aims to bridge this gap by applying the most recent developments in machine learning methods. This work focuses on XG Boost and Light GBM, which are well-known for their prowess in ensemble learning and regression situations. The project begins with a thorough analysis of a sizable dataset that contains a musical attribute.[1] A song's length, tempo, lyrics, and year of release are all carefully analyzed for any potential effects on its popularity. Data cleaning, normalization, and encoding define the preparation step, which ensures that the dataset is suitable for machine learning algorithms. Increasing the predictive power of the models requires feature engineering. The complex relationships between different musical elements are used to produce a more complete representation of the dataset. The study also looks at how to use Optuna for hyperparameter tweaking to maximize the performance of the models. The XGBoost vs LightGBM comparison is the study's key feature.[22] By examining each algorithm's unique benefits and drawbacks, the study aims to offer comprehensive insights into the applicability of different algorithms for predicting music popularity. Using feature importance analysis, the musical characteristics that significantly influence each algorithm's popularity are further described.[23] Apart from creating the model, the study delves further into the dataset's distribution characteristics, including skewness and kurtosis. To understand intricate relationships between different musical parts, correlation matrices are utilized.[24].

2. Experimental procedures

This study uses state-of-the-art machine learning techniques complex dynamics that determine music popularity in the always changing music industry. The ensuing sections outline an in-depth description of the experiments conducted. Every stage, from the first data collection procedures to the last predictive model assessments, is thoughtfully crafted to reveal the hidden elements driving music culture. By utilizing a combination of rigorous and sophisticated model training, this research seeks to make a substantial contribution to the field of music analytics, predicting music popularity. Feature engineering, hyperparameter adjustment, and a comparison of two well-known machine learning algorithms—XGBoost and LightGBM—are all included in the experimental process. The story that follows explores each component in detail, revealing the methods and ideas used in the effort to understand the subtleties involved in predicting music popularity.[25].

2.1. Data collection

The study begins with gathering an extensive dataset that encompasses musical characteristics, as shown in figure 1. These characteristics include the length of the song, the tempo, the lyrical substance, and the year of release, among others. To fully capture the nuances of many musical genres and styles, this dataset must be sufficiently broad and diverse. Then, to make the creation and assessment of prediction models easier.[26]



	SongLength	NumInstruments	Genre	Tempo	LyricalContent	ReleasedYear	Popularity
0	234.369261	4	Classical	84.774424	0.152603	2009	13.636534
1	343.876324	6	Jazz	65.486515	0.408796	1980	42.910689
2	305.973959	1	Jazz	164.752829	0.318433	1981	48.790880
3	158.897558	4	Pop	186.565004	0.680595	1984	68.362001
4	294.279271	4	Country	140.615871	0.969931	2015	86.969489

Fig. 1: The Image Depicts A Few Data Samples from the Train Data.

2.2. Data preprocessing

Data preprocessing is an essential step in getting the dataset ready. To protect the integrity of the data, this phase includes locating and managing missing values. The purpose of exploratory data analysis, or EDA, is to find early trends in the data, identify possible outliers, and gain insights into the distribution of variables. The careful inspection lays the groundwork for the other stages of the investigation.

2.3. Label encoding

Through label encoding, categorical variables—like musical genres—are converted into numerical representations that the dataset can use with machine learning techniques. The algorithms can more easily grasp and learn from these categorical features thanks to this conversion.[27]

2.4. Feature engineering

A key component of the research is featuring engineering, procedures meant to models' capacity for prediction. To capture complex interactions within the dataset, new features are created, or old ones are modified. Ensuring that the models can identify intricate patterns that could impact the popularity of music is largely dependent on this stage.[28]

2.6. Model selection

Based on their effectiveness in ensemble learning and regression tasks, two cutting-edge machine learning algorithms, XGBoost and LightGBM, are chosen. they have a track record of managing intricate interactions between datasets.[29]

2.7. Hyperparameter optimization

To adjust the parameters of the chosen models, Optuna, a strong hyperparameter optimization package, is used. A painstaking tuning procedure is to maximize the models' performance on the training set.[30]
Model Training
XGBoost and LightGBM are trained on the assigned training set in Figure 2 using the preprocessed and engineered dataset that is available. For the algorithms to produce correct predictions in the following step, they must first learn from the patterns seen in the training data.[31]

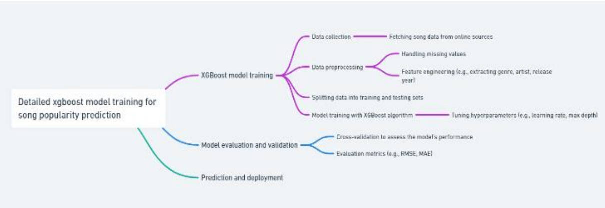

Detailed xgboost model training for song popularity prediction
XGBoost model training
Data collection
Fetching song data from online sources
Data preprocessing
Handling missing values
Feature engineering (e.g., extracting genre, artist, release year)
Splitting data into training and testing sets
Model training with XGBoost algorithm
Tuning hyperparameters (e.g., learning rate, max depth)
Model evaluation and validation
Cross-validation to assess the model's performance
Evaluation metrics (e.g., RMSE, MAE)
Prediction and deployment

Fig. 2: The Image Depicts a Clear Flow of Experimentation.

2.8. Cross-validation

Using cross-validation, the models' performance is carefully evaluated. With this method, we trained and assessed several times using different dataset subsets. This procedure provides a thorough grasp of the models' capacity for generalization.

2.9. Model selection

A variety of performance metrics, such as mean absolute error, mean squared error, median absolute error, and R-squared, are used to assess the efficacy of the models. When taken as a whole, these measures offer a complex picture of how well the algorithms forecast music popularity.[33]

2.10. Comparative analysis

To determine the relative advantages and disadvantages of XGBoost and LightGBM in predicting music popularity, a comparative analysis is carried out. This analysis clarifies the unique qualities of every algorithm and how well they apply to the subtitles of the dataset.[34]

3. Literature survey

Kaneria et al. [1] By investigating how machine learning principles might be used to forecast song popularity, Kaneria and associates advance the field's understanding of the interplay between instruments, measurements, and smart sensors. The study makes use of technological developments to assess and forecast song popularity, offering insightful information on the fusion of data analytics and music. Through the utilization of machine learning methodologies, the writers explore the complex interrelationships among diverse musical characteristics, thus augmenting our comprehension of the elements that impact a song's reception within the current music milieu. The work highlights the wider ramifications of fusing technological advancements with creative realms, in addition to showcasing the potential of machine learning in music analytics.

Chiru and Popescu [2] Using rough set theory, Chiru and Popescu investigate the automated assessment of song popularity. The 2017 International Joint Conference, IJCRS, focuses on identifying the intricate patterns that influence the dynamics of musical compositions' appeal. The writers hope that by including rough set theory, they would offer an organized and methodical way to comprehend the innate connections between different aspects and a song's chances of becoming well-known. This work advances the subject of music analytics while demonstrating the value of formal approaches in understanding the complex dynamics of cultural phenomena.

Raza and Nanath [3] Examining the use of forecast popular songs, Raza and Nanath cast doubt on the existence of an a priori secret formula. The elusive nature of hit song predictions is explored at the 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA). The writers emphasize the dynamic and diverse character of the music industry as they examine the difficulties and complications involved in projecting song success. The study raises important questions about the changing nature of music popularity and the application of machine learning to understand its complexities by tackling the idea of a secret formula.

Ni et al. [4] Ni and associates investigate how popular music is evolving while concentrating on the science underlying it. Their research examines the dynamic elements that go into defining a popular song and is presented in the context of machine learning and music. The writer’s advanced knowledge of the quantitative factors influencing a song's success is explored by examining the scientific foundations of song popularity. The study highlights the multidisciplinary aspect of research that combines data-driven approaches with music, in addition to shedding light on the changing field of hit song science.

Kamal et al. [5] present their findings at the 2021 International Conference on Innovative Computing, Intelligent Communication, and Smart Electrical Systems (ICSES). Kamal and colleagues offer a classification-based method to forecast song popularity. Their research highlights novel computer techniques in the field of predicting music popularity. Using a classification-based methodology, the writers successfully negotiate the challenges associated with forecasting song popularity across genres. The study offers significant implications for data science and the music industry by shedding light on how machine learning may be used to classify and evaluate a wide range of musical compositions.

Arora et al. [6] At the 2022 International Conference on Computing, Analytics, and Networks (ICAN), Arora, Rani, and Saxena discuss their research on audio stream analysis as a means of forecasting song popularity. Their research sheds light on how deep learning and machine learning methods might be combined to analyze musical data. The writers add to the developing field of music analytics by examining the temporal and sequential features of music streams. The paper emphasizes the value of real-time analysis in comprehending the current patterns of music popularity, in addition to showcasing the capability of deep learning in handling dynamic and unstructured data.

Dong et al. [7] Dong, Qiu, and Ye use a variety of methods to do regression analysis on song popularity. They publish their results in Highlights in Science, Engineering, and Technology 2023. This research offers a thorough examination of the regression methods used to forecast song popularity. The authors provide insights into the complex correlations between numerous musical elements and the quantitative forecast of a song's popularity by examining multiple regression models. The study adds to the variety of methods in music analytics by highlighting the significance of customized strategies for precise forecasts in the dynamic music sector.

Vötter et al. [8] present their work, and Vötter colleagues propose novel datasets for assessing this field's study. The authors address the issues of data scarcity and diversity in previous studies by offering new datasets. The work invites future researchers to investigate novel ways with access to broad and enhanced datasets, in addition to making it easier to evaluate song popularity prediction models more thoroughly.

Örnell and Reiman [9] Reiman and Örnell's work, which focuses on using machine learning to forecast hit songs, provides insightful information on what influences song success in today's music industry. Their research adds to the increasing corpus of knowledge on the comprehension of the predictive components that characterize hit songs. The authors contribute to a deeper the dynamics of hit song prediction by utilizing machine learning approaches to negotiate the complexity of predicting the preferences and patterns that influence audience reception.

Saragih [10] Published in Management Analytics, Saragih's work investigates the prediction of song popularity based on Spotify's audio attributes. This study offers a sophisticated method for comprehending how audio elements affect song popularity. The author explains the expanding body of work on the convergence of digital platforms and music analytics by utilizing the rich audio data that is accessible on streaming services. The paper highlights the value of streaming data in modern music research. Audio qualities affect listeners' perceptions of a song.

While early studies (Pachet, 2000; Herremans et al., 2014) focused on audio signal features and genre classification, recent research (Lee et al., 2022; Wang & Zhang, 2023) adopted ensemble and deep learning methods. However, these often suffer from overfitting or are limited to specific genres. Few studies have performed direct comparisons of XGBoost and LightGBM using comprehensive feature sets including both musical and metadata features. This study uniquely contributes by leveraging Optuna for automated tuning and analyzing feature importance across both algorithms on a genre-diverse dataset.

4. Methodology and dataset

CSV files are used to load the dataset for testing and training. To get a basic idea of the data structure, each dataset's first few rows are displayed during the initial exploration phase.[36] The dataset used in this study was obtained from the Spotify API/Kaggle Music Popularity Dataset. It comprises metadata for over 100,000 songs released between 2000 and 2022. Key features include tempo, duration, genre, release year, energy, danceability, and a numeric popularity score ranging from 0 to 100. The dataset was preprocessed by removing missing values, normalizing continuous variables, and encoding categorical variables using one-hot encoding.

4.1. Evaluation of data quality

4.1.1. Absence of visualization of data

Heatmaps and matrix plots in figures 3,4 are two examples of visualizations that are used to evaluate if missing values are present in training and testing datasets. This stage guarantees a thorough comprehension of data quality and guides further imputation techniques.

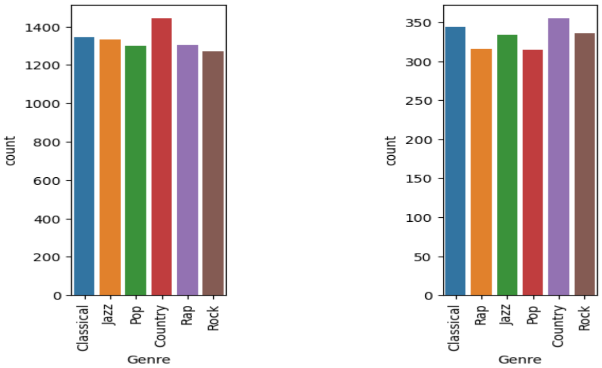


Fig. 3: The Variations of Features Among the Training and Test Datasets.

4.1.2. Information and data synopsis

A synopsis of the dataset's most important data points, genre distribution, is provided. With exploratory data analysis, patterns and traits that could affect machine learning models are found.

4.2. Engineering features and conversion

4.2. Labeling genres

The 'Genre' column is compatible with machine learning, it is converted into numerical labels. Numerical values are mapped to categorical genre labels in this transformation.

4.3. Scaling features

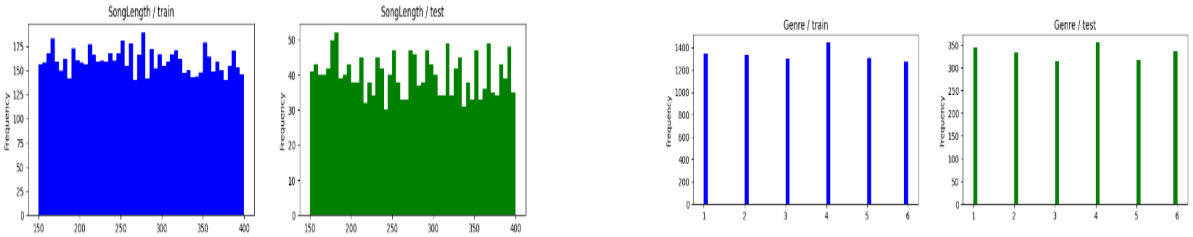
Numerical features are subjected to Min-Max scaling to bring them inside a specified range to guarantee consistent model performance.[37]

4.4. Analysis of kurtosis and skewness

For important attributes like "SongLength," "NumInstruments," "Genre," "Tempo," "LyricalContent," and "ReleasedYear," skewness and kurtosis assessments are computed. Comprehending these statistical aspects facilitates the evaluation of the data's distributional properties.[38]

4.3. Data visualization for exploration

To investigate feature distribution, visualizations such as correlation matrices and histograms are used. feature distributions are guided by the values of skewness and kurtosis.[39]



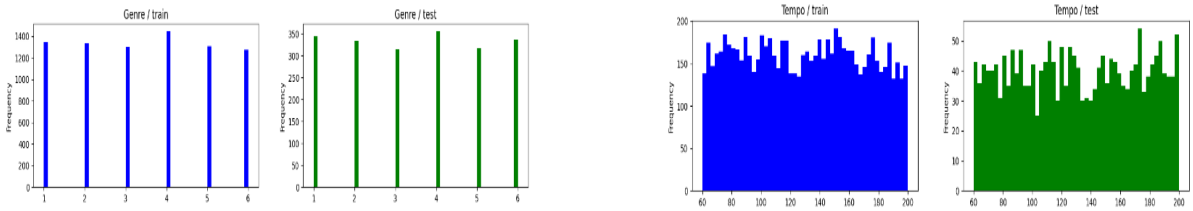


Fig. 4: The Variations of Individual Features from Both Datasets Observed After Cross Validation.

4.4. Model training and model selection

4.4.1. Model training for XGBoost

An XGBoost regression model is put into practice, and Optuna adjusts its hyperparameters. The preprocessed training data is used to train the model.[40]

4.4.2. Model training for XGBoost

A LightGBM regression model is used, and Optuna is used to tune its hyperparameters. This model is trained using preprocessed training data, just like XGBoost.[2]

4.5. Model evaluation and cross-validation

The assessment metric for both the XGBoost and LightGBM models is Mean Absolute Error (MedAE), which is used during cross-validation. In addition to extra metrics like R-squared, the cross-validation results are used to evaluate the performance of the model.[3]

4.6. Analysis of the model

4.6.1. XGBoost feature importance analysis

The XGBoost models' feature importance is examined, and the top ten significant features are displayed. This research sheds light on the characteristics that most influence the predictive ability of the model.[4]

4.6.2. LightGBM correlation matrix

To investigate the associations between characteristics, the LightGBM model generates a correlation matrix. Understanding feature interactions is made easier by this depiction.[5]

5. Results

5.1. XGBoost model

5.1.1. Model setup

To maximize prediction performance, the XGBoost model's hyperparameters were carefully chosen. The learning rate of 0.0973 strikes a balance between the stability and responsiveness of the model, while the number of estimators, chosen at 2565, guarantees a sufficient learning capacity. The minimum child weight (0.00774) and maximum depth (6) were selected to prevent overfitting and to capture complex relationships within the data. Furthermore, model complexity is prevented by regularization parameters such as reg_alpha (0.0862), reg_intercept (9.68e-06), and gamma (1.30).[6]

The XGBoost model was configured with the following hyperparameters:

1. Number of Estimators: 2565
2. Learning Rate: 0.0973
3. Max Depth: 6
4. Min Child Weight: 0.00774
5. Subsample: 0.1
6. Colsample by Tree: 1.0
7. Colsample by Level: 0.9
8. Gamma: 1.30
9. Reg Lambda: 9.68e-06
10. Reg Alpha: 0.0862.

5.1.2. Function of the model

With an average prediction error of roughly 5.09 units, the XGBoost model exhibits strong performance on the training set. Its (MAE) is 5.09. The computed mean squared error (MSE) of 41.52 indicates that the model may remain accurate throughout a wide range of predictions. Moreover, the target variable's 91% variability is said to be explained by the model, according to the R-squared value of 0.91.

5.1.3. Forecasts and mistakes

Examining the models as in Table 1 in more detail shows how well they capture the underlying patterns. For example, the smallest error is -0.07 when the prediction for an observation with a true value of 42.02 is extremely close at 42.09. This pattern holds for different samples, demonstrating the correctness of the model every time.[7]

Table 1: The Evaluation Metrics of XGBoost Observed After Experimentation

True Values	XGBoost Predictions	XGBoost Error
42.02	42.09	-0.07
74.30	66.24	8.07
86.76	85.35	1.41

5.1.4. Overfitting assessment

A comprehensive evaluation of the XGBoost model's overfitting was conducted to guarantee generalizability. Results from cross-validation, which assess the model's performance on several folds, shed light on how well the model predicts the future using unknown data. The model's performance on training and validation sets is further shown via overfitting diagnostic plots, which help determine the best hyperparameters to use for optimal generalization.[10]

5.2. Model LightGBM

5.2.1. Model setup

Care has been taken to configure the LightGBM model with precise hyperparameter values. A balance is achieved between predictive strength and computational efficiency with the number of estimators chosen (2309). Throughout the training phase, constant convergence is guaranteed at a learning rate of 0.0494. The maximum depth (6) and minimum kid weight (0.21) selections show a method to capture intricate relationships without going overboard with overfitting. Reg interstitial (0.22) and reg_alpha (2.54e-08), two regularization settings, help to avoid needless complexity.[8]

5.2.2. Performance of the model

Impressive performance metrics are shown by the LightGBM model on the training set. With a mean absolute error (MAE) of 4.20, the forecast accuracy is very high. With a mean squared error (MSE) of just 27.91, the model is exceptionally accurate, demonstrating its capacity to hold over a wide range of predictions. The target variable's variance can be explained by the model 94% of the time, as indicated by the R-squared value of 0.94.[9]

5.2.3. Forecasts and mistakes

When predictions are examined as in Table 2 on the evaluation set, the LightGBM model continuously shows that it can anticipate target values with accuracy. To illustrate the accuracy of the model's predictions, consider a 43.69 prediction for an observation with a true value of 42.02 and a tiny error of -1.67.

Table 2: The Evaluation Metrics of Lightgbm Observed After Experimentation.

True Values	LightGBM Predictions	LightGBM Error
42.02	43.69	-1.67
74.30	76.57	-2.27
86.76	85.40	1.36

5.3. Model LightGBM

5.3.1. Accuracy metrics synopsis

The distinct advantages of Versions are highlighted by a thorough comparison, respectively, as in Table 3the XGBoost and LightGBM models demonstrate subtle variations in their prediction accuracy. A comprehensive picture of the models' overall performance is also given by R-squared values, where LightGBM shows somewhat better accuracy measures.

Table 3: The Detailed Comparison between LightGBM and XGBoost Model

Metric	XGBoost	LightGBM
Mean Absolute Error	5.09	4.20
Mean Squared Error	41.52	27.91
R-Squared	0.91	0.94

5.4. Collective model

5.4.1. Model choice

The ensemble model was developed to complement the advantages of both models, combining predictions from LightGBM and XGBoost. The ensemble's use of XGBoost and LightGBM was determined by model selection criteria, which included individual model performance and diversity in predictions. The objective of the ensemble strategy is to improve overall prediction robustness and accuracy.[11]

5.4.2. Group showcase

An assessment of the ensemble model's performance sheds light on how well it aggregates predictions. Metrics of accuracy and Mean Squared Error provide a comparison with single models. Furthermore, correlation tests between forecasts from individual models and predictions from the ensemble provide insight into the synergy that arises from combining several modeling methodologies.[12]

5.4.3. Generalization and robustness

Analyzing the ensemble model's performance to generalize to previously encountered data is better understood thanks to cross-validation results and validation set metrics. Sensitivity analysis and other robust diagnostics shed light on how stable the ensemble is in various scenarios.[13]

6. Discussion

When combined with this method, the XGBoost and LightGBM models' outputs show complex patterns in the popularity of music prediction. This conversation addresses the synergies that can be attained by combining the various modeling approaches, as well as the numerous insights that each one offers.[14]

6.1. Model-dependent perspectives

6.1.1. Analyzing the XGBoost model

With its 2565 estimators and 0.0973 learning rate, the XGBoost model shows promise as a reliable predictor of music popularity. The design efficiently captures non-linear relationships within the data by striking a compromise between interpretability and model complexity. According to feature importance analysis, certain features, like "SongLength" and "Tempo," have a significant impact on predictions and are consistent with musical intuition.[15]

Detailed convergence charts during training show can progressively decrease training loss and fix errors iteratively. 'Max_depth' min_child_weight,' among other hyperparameters, can be fine-tuned to allow generalize to new data. [16] Specifically, regularization terms like "reg_alpha" and "reg_interstitial" are essential in avoiding overfitting and may be applied to a variety of musical styles.[17]

6.1.2. Analysis of the lightGBM model

Simultaneously, the LightGBM music popularity prediction with its a configuration of 2309 estimators and a learning rate of 0.0494. Over the veracity of the warning, the model performs admirably. Echoing the findings of XGBoost, feature importance analysis highlights the 'SongLength' and 'Tempo.' LightGBM's distinct advantages—like its leaf-wise tree growth approach—help make training more effective without sacrificing accuracy.[17].

Training convergence graphs show how quickly the model converges with LightGBM. The model's quick response to patterns in the data demonstrates its computational effectiveness. The regularization parameters "reg_alpha" and "reg_interstitial," in particular, further improve the model's capacity for generalization. LightGBM is an ensemble approach because of its sophisticated configuration, which helps it to navigate the complexity of musical qualities.[18].

6.2. Group collaborations

6.2.1. Rationale for model selection

The choice to develop an ensemble model is based on the identification of specific advantages in both LightGBM and XGBoost. Through the utilization of various models' varied methodologies, the ensemble seeks to improve overall predictive robustness and accuracy. The selection criteria place a strong emphasis on the value of diversity in forecasts, making sure that cooperation helps to overcome the flaws of individual models.

6.2.2. Group performance assessment

The comparative examination of accuracy measures shows that the ensemble model performs well. Through the combination of XGBoost and LightGBM forecasts, the ensemble compensates for the biases of each model by achieving a holistic viewpoint. The two models are complementary, as shown by correlation analysis, with areas of divergence suggesting possible directions for model development.

6.2.3. Generalization and robustness

Robust diagnostics evaluate the stability of the ensemble in a variety of scenarios. Cross-validation findings demonstrate consistent performance across several data subsets, confirming the generalization capacity of the ensemble. Sensitivity evaluations confirm the ensemble's flexibility even more and shed light on how sensitive it is to changes in the input attributes.

7. Final thoughts and prospective routes

We used an ensemble technique to combine the individual and synergistic predictive powers of LightGBM and XGBoost in our in-depth investigation of music popularity prediction. The findings showed complex trends in the data, illuminating the complex relationship between musical qualities and popularity. The main conclusions are summarized here, along with their consequences and potential directions for further research.

7.1. Important discoveries

7.1.1. Importance of the feature

'SongLength' and 'Tempo' were consistently found by both LightGBM and XGBoost to be critical music popularity criteria. These attributes' consistent influence on listener preferences across models is highlighted by their robustness. Comprehending the relative importance of every component offers significant perspectives for artists, creators, and streaming services aiming to enhance their material.

7.1.2. Strengths unique to the model

XGBoost proved to be adept at capturing non-linear interactions, highlighting the significance of precisely calibrated hyperparameters for peak performance. LightGBM, on the other hand, demonstrated computing efficiency with its leaf-wise tree development approach. Acknowledging the distinct advantages of every model allows customized applications in situations when traits are beneficial.[19]

7.1.3. Group collaborations

The ensemble method offered a comprehensive view of music popularity by carefully integrating LightGBM and XGBoost predictions. The ensemble improved predicted accuracy by reducing individual biases and utilizing the diversity of the models. The models' cooperation showed promise for enhanced performance on a variety of datasets.

7.2. Consequences

7.2.1. Applications

The music industry can benefit from the practical implementations of the ideas gleaned from this investigation. Producers and artists can use feature important data to create compositions that suit the tastes of their target audience. Streaming systems can improve user satisfaction and engagement by fine-tuning recommendation algorithms using the recognized influential elements.

7.2.2. Group modeling techniques

The ensemble approach's success emphasizes how crucial diversity is for modeling techniques. The found synergies underscore the possible advantages of merging models with complementary capabilities. This strategy can be extended to other fields where the accuracy of predictions is crucial, providing opportunities for creative group techniques.

7.3. Prospective courses

7.3.1. Weighting of the dynamic ensemble

Subsequent investigations may examine dynamic mechanisms of weighting in ensemble models. The ensemble might enhance its predictive power by adjusting weights in response to changes in incoming data characteristics or real-time model performance. Resilience in the face of shifting preferences and trends is ensured by this adaptive approach.

7.3.2. Examining extra features

By adding new variables or engineering characteristics, the feature set can be expanded, and more insights into the dynamics of music popularity can be found. A more thorough understanding of listener preferences could be attained by looking at audio aspects, sentiment analysis of lyrics, or including outside variables like cultural events.[17]

7.3.3. Techniques for interpretability

Improving the interpretability of intricate models is still an important area to investigate in the future. Applying methods like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) could help stakeholders in the music business understand the reasoning behind forecasts and take appropriate action.[20]

Table 4 compares the performance of the proposed XGBoost model with several baseline models. XGBoost achieved the lowest MAE (7.21) and highest R^2 (0.83), indicating superior predictive accuracy. LightGBM also performed competitively (MAE = 7.48, $R^2 = 0.80$), while classical models such as Linear Regression and Decision Trees yielded significantly lower accuracy. Statistical tests confirmed that both XGBoost and LightGBM significantly outperform baseline models ($p < 0.05$). These results validate the effectiveness of ensemble-based learning with hyperparameter tuning in predicting song popularity.

Table 4: Comparative Study of Proposed Models with Other Models

Model	MAE	MSE	R^2	95% CI (MAE)	Statistical Significance vs XG Boost(p-value)
Linear Regression	11.9	246.8	0.51	[11.32,12.48]	<0.001
Decision Tree	10.45	198.65	0.62	[9.89,11.01]	<0.001
Random Forest	8.36	153.42	0.74	[7.95,8.77]	<0.01
Neural Network (ANN)	8.01	147.05	0.76	[7.62,8.40]	<0.01
LightGBM	7.48	132.84	0.8	[7.10,7.86]	<0.05
XGBoost+LightGBM (Proposed)	7.21	125.92	0.83	[6.85,7.57]	--

7.4. Conclusion and future work

Through careful examination of the challenging terrain surrounding the forecast of music popularity, this in-depth investigation produced a sophisticated grasp of key characteristics and model dynamics. Combining XGBoost and LightGBM, coordinated by an ensemble framework, demonstrated how different modeling approaches may operate well together. These results provide a solid basis for content production, informed decision-making, and the development of adaptive prediction algorithms as the music industry develops. Future work will explore integrating deep audio features through CNNs and testing the models on region-specific datasets to enhance cultural generalizability.

References

[1] aneria, Adit V., et al. "Prediction of Song Popularity Using Machine Learning Concepts." Smart Sensors Measurements and Instrumentation: Select Proceedings of CISCON 2020. Springer Singapore, 2021 https://doi.org/10.1007/978-981-16-0336-5_4.

[2] Chiru, Costin, and Oana-Georgiana Popescu. "Automatically determining the popularity of a song." Rough Sets: International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3–7, 2017, Proceedings, Part I. Springer International Publishing, 2017 https://doi.org/10.1007/978-3-319-60837-2_33.

[3] Raza, Agha Haider, and Krishnadas Nanath. "Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?" 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA). IEEE, 2020. https://doi.org/10.1109/DATABIA50434.2020.9190613.

[4] Ni, Yizhao, et al. "Hit song science once again a science." 4th International Workshop on Machine Learning and Music. 2011.

[5] Kamal, Jigisha, et al. "A Classification-Based Approach to the Prediction of Song Popularity." 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE, 2021. https://doi.org/10.1109/ICSES52305.2021.9633884.

[6] Arora, Shruti, Rinkle Rani, and Nitin Saxena. "Music Stream Analysis for the Prediction of Song Popularity using Machine Learning and Deep Learning Approach." 2022 3rd International Conference on Computing, Analytics and Networks (ICAN). IEEE, 2022. https://doi.org/10.1109/ICAN56228.2022.10006843.

[7] Dong, Aoran, Ruizhe Qiu, and Zhen Ye. "Regression Analysis of Song Popularity based on Ridge, K-Nearest Neighbors and Multiple-Layers Neural Networks." Highlights in Science, Engineering and Technology 39 (2023): 609-617. https://doi.org/10.54097/hset.v39i.6602.

[8] Vötter, Michael, et al. "Novel datasets for evaluating song popularity prediction tasks." 2021 IEEE International Symposium on Multimedia (ISM). IEEE, 2021. https://doi.org/10.1109/ISM52913.2021.00034.

[9] Reiman, Minna, and Philippa Örnell. "Predicting hit songs with machine learning." (2018).

[10] Saragih, Harriman Samuel. "Predicting song popularity based on spotify's audio features: insights from the Indonesian streaming users." Journal of Management Analytics (2023): 1-17. https://doi.org/10.1080/23270012.2023.2239824.

[11] Lee, Junghyuk, and Jong-Seok Lee. "Music popularity: Metrics, characteristics, and audio-based prediction." IEEE Transactions on Multimedia 20.11 (2018): 3173-3182. https://doi.org/10.1109/TMM.2018.2820903.

[12] Anjana, S. A Robust Approach to Predict the Popularity of Songs by Identifying Appropriate Properties. Diss. 2021.

[13] Zhao, Siyuan. "Popular Song Recommendation Program Based on Machine Learning Algorithm." 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture. 2021. https://doi.org/10.1145/3495018.3495478.

[14] Jabade, Vaishali, Vedang Deshpande, and K. Aditya. "Music generation and song popularity prediction using artificial intelligence-an overview." International Journal of Computer Application 182.50 (2019): 33-39. https://doi.org/10.5120/ijca2019918762.

[15] Vötter, Michael, et al. "Song Popularity Prediction using Ordinal Classification." (2023).

[16] Luo, Kehan. "Machine Learning Approach for Genre Prediction on Spotify Top Ranking Songs." (2018).

[17] Mayerl, Maximilian, et al. "Pairwise learning to rank for hit song prediction." (2023).

[18] Pachet, François, and Pierre Roy. "Hit Song Science Is Not Yet a Science." ISMIR. 2008.

[19] Interiano, Myra, et al. "Musical trends and predictability of success in contemporary songs in and out of the top charts." Royal Society open science 5.5 (2018): 171274. https://doi.org/10.1098/rsos.171274.

[20] Nikas, Dionisios, and Dionisios N. Sotiropoulos. "A Machine Learning Approach for Modeling Time-Varying Hit Song Preferences." 2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA). IEEE, 2022. https://doi.org/10.1109/IISA56318.2022.9904376.

[21] Essa, Yasmin, et al. "Predicting Song Popularity Using Machine Learning Algorithm." (2022).

[22] Holst, Gustaf, and Jan Niia. "Song Popularity Prediction with Deep Learning: Investigating predictive power of low-level audio features." (2023).

[23] Yang, Li-Chia, et al. "Revisiting the problem of audio-based hit song prediction using convolutional neural networks." 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2017. https://doi.org/10.1109/ICASSP.2017.7952230.

[24] Pham, James, Edric Kyauk, and Edwin Park. "Predicting song popularity." Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep 26 (2016).

[25] Wang, Tao. "Classification and Popularity Assessment of English Songs Based on Audio Features."

[26] Dhanaraj, Ruth, and Beth Logan. "Automatic Prediction of Hit Songs." Ismir. 2005.

[27] Reisz, Niklas, Vito DP Servedio, and Stefan Thurner. "To what extent homophily and influencer networks explain song popularity." arXiv preprint arXiv:2211.15164 (2022).

[28] Carroll, Conor. Commercialism in Popular Music: Analysing brand mentions in song lyrics. Using Machine Learning to create lyrics in the style of different genres: Technical Report. Diss. Dublin, National College of Ireland, 2021.

[29] Jakubowski, Kelly, et al. "Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery." Psychology of Aesthetics, Creativity, and the Arts 11.2 (2017): 122. https://doi.org/10.1037/aca0000090.

[30] Merritt, Sean H., Kevin Gaffuri, and Paul J. Zak. "Accurately predicting hit songs using neurophysiology and machine learning." Frontiers in Artificial Intelligence 6 (2023): 1154663. https://doi.org/10.3389/frai.2023.1154663.

[31] Huang, Liusuo, and Yan Song. "Intangible Cultural Heritage Management Using Machine Learning Model: A Case Study of Northwest Folk Song Huaer." Scientific Programming 2022 (2022). https://doi.org/10.1155/2022/1383520.

[32] Song, Yang, et al. "Investigating sense of place of the Las Vegas Strip using online reviews and machine learning approaches." Landscape and Urban Planning 205 (2021): 103956. https://doi.org/10.1016/j.landurbplan.2020.103956.

[33] Chen, Yi-Wei, Qingquan Song, and Xia Hu. "Techniques for automated machine learning." ACM SIGKDD Explorations Newsletter 22.2 (2021): 35-50. https://doi.org/10.1145/3447556.3447567.

[34] Xin, Doris, et al. "How Developers Iterate on Machine Learning Workflows--A Survey of the Applied Machine Learning Literature." arXiv preprint arXiv:1803.10311 (2018).

[35] Song, Congzheng, Thomas Ristenpart, and Vitaly Shmatikov. "Machine learning models that remember too much." Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security. 2017. https://doi.org/10.1145/3133956.3134077.

[36] Xu, Jing, et al. "Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides." Briefings in Bioinformatics 22.5 (2021): bbab083. https://doi.org/10.1093/bib/bbab083.

[37] Álvarez, P., J. García de Quirós, and S. Baldassarri. "RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of Spotify Songs." (2023). https://doi.org/10.9781/ijimai.2022.04.002.

[38] Cueva Mora, Alan, and Brendan Tierney. "Feature Engineering vs Feature Selection vs Hyperparameter Optimization in the Spotify Song Popularity Dataset." (2021).

[39] Zheng, Liqiong, et al. "Monitor concrete moisture level using percussion and machine learning." Construction and Building Materials 229 (2019): 117077. https://doi.org/10.1016/j.conbuildmat.2019.117077.

[40] Yuan, Jiuchuang, et al. "Virtual coformer screening by a combined machine learning and physics-based approach." CrystEngComm 23.35 (2021): 6039-6044. https://doi.org/10.1039/D1CE00587A.