Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241655

Predictive analytics with music: Advancing tree-based models for song rating prediction

Jiaxuan Xu

Columbia University, 444 Warren ST, Apt 3556, Jersey City, NJ 07302

13383578008@163.com

Abstract. This paper presents an in-depth exploration of predictive analytics applied to a dataset of 19,485 songs from the Kaggle Predictive Analysis Competition (PAC), with the objective of forecasting song ratings based on auditory features. The study employs a range of tree-based models, including regression tree, bagging, and random forest, and confronts various data preprocessing challenges, particularly in handling missing data and incorporating genre as a significant predictive feature. Through the creation of dummy variables for genre classification and careful model selection, the research demonstrates an enhanced approach to predictive accuracy. The effectiveness of these models is rigorously evaluated using test Root Mean Square Error (RMSE), providing valuable insights into their predictive performance. This paper contributes to the field of music analytics by offering a comprehensive analysis of tree-based predictive models and their application in the nuanced task of song rating prediction, highlighting the importance of methodological refinement in predictive analytics.

Keywords: Predictive Analysis, Auditory Features, Machine Learning.

1. Introduction

In the expansive domain of machine learning, the application of tree-based models has demonstrated significant versatility and effectiveness across diverse fields, suggesting a promising approach for predictive analytics in music. Previous research has validated the effectiveness of these methods in survey research, laying a groundwork that underscores their potential applicability in music analytics [1]. This is further supported by a study where tree-based models were adeptly used to forecast stock market trends [2], demonstrating their proficiency in managing complex data sets, akin to the complexity of auditory features in music.

While machine learning has been extensively used in music analysis, the primary focus has been on predicting musical features rather than leveraging these features to predict a song's popularity or ratings. This study seeks to address this gap. Earlier work involved using machine learning to dissect compositional patterns in music [3], with a focus on the music's inherent characteristics. Another study developed a tree-structured model for written music [4], highlighting the structural aspects of compositions. Although these methods provide valuable insights into music's intrinsic features, they do not directly predict a song's reception based on these characteristics. Additionally, a different approach utilized classification models for music similarity estimation [5], a strategy aligned with this study's goals but primarily confined to analyzing music features, rather than their impact on audience response.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/52/20241655

This study, therefore, ventures into a relatively unexplored area of using machine learning, particularly tree-based models, to predict song ratings based on auditory features. This approach not only contributes to the academic understanding of how auditory features influence public preference but also has practical implications in the music industry, particularly in refining music recommendation systems. By shifting the focus from the analysis of music features to their predictive power regarding audience reception, this research provides a novel perspective in the field of music analytics. This research aims to explore the efficacy of various tree-based predictive models in accurately forecasting song ratings, a task pivotal for tailoring music recommendation systems and understanding listener preferences. By assessing models like regression trees, bagging, and random forests, the study offers a nuanced understanding of model performance, particularly focusing on genre as a key predictive feature. The significance of this research lies in its potential to refine predictive models in music analytics, contributing to both theoretical understanding and practical applications in digital music platforms.

2. Exploratory Data Analysis (EDA)

In examining the "analysisData" dataset, this study identified a range of features characterizing each song, including performer, song title, track duration, track explicitness, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time signature, and genre. To ascertain a model apt for predicting song ratings, an analysis of the distribution of relationships between independent variables and ratings was conducted. Utilizing the 'ggpairs' function, scatterplot matrices were generated for the numerical variables within "analysisData". Particularly noteworthy are the scatterplots and trend lines in the final row, which illustrate the relationship between each variable and the corresponding song rating values.

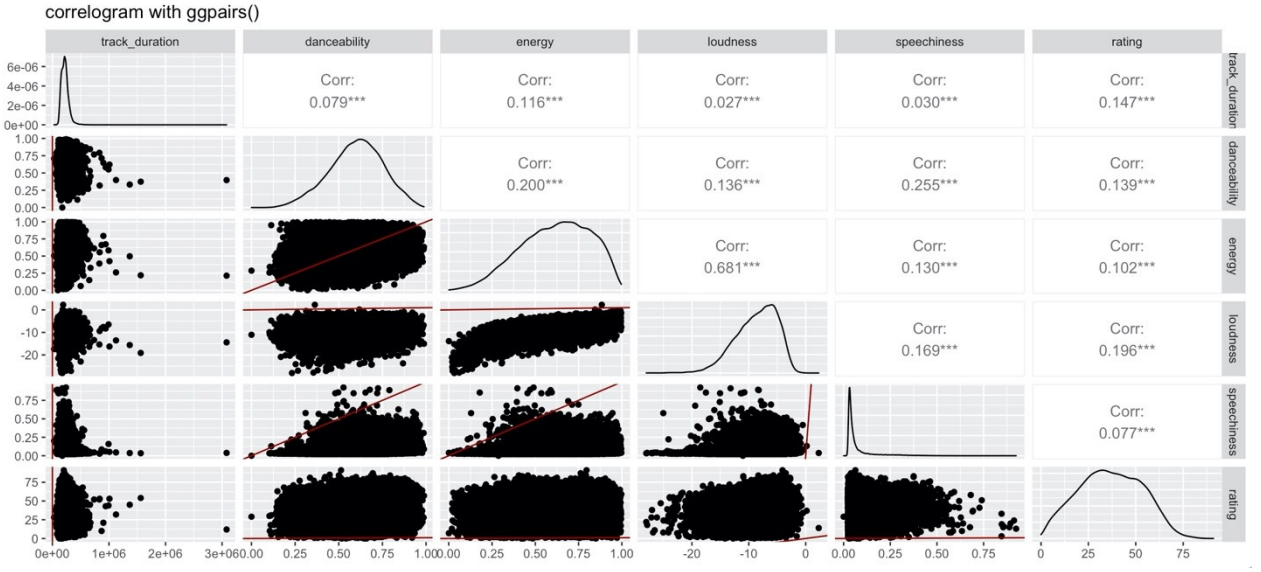

Figure 1. scatterplots matrix 1

DOI: 10.54254/2755-2721/52/20241655

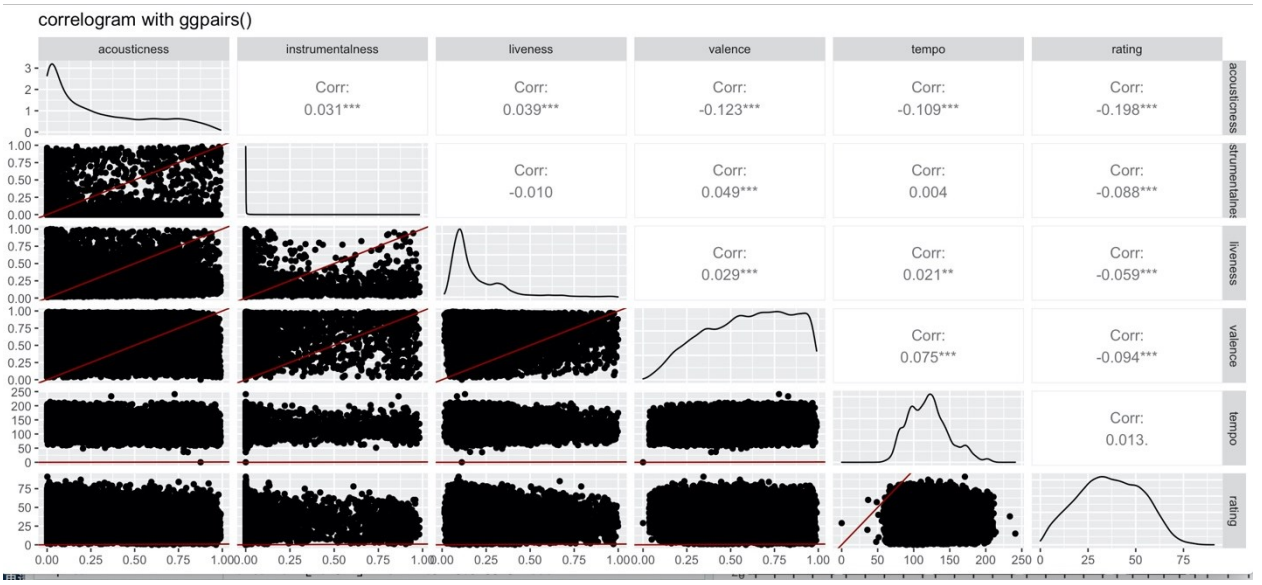

Figure 2. scatterplots matrix 2

The scatterplot matrix analysis of the variables in the dataset revealed a notable dispersion of the numerical variables relative to the trend lines, suggesting that a linear regression model may not be optimal for this predictive task. Observing the spherical distribution of variables, the study pivoted towards various predictive tree models. These models are advantageous as they stratify the predictor space into multiple simple regions, with predictions deriving from a summary statistic of the outcomes in each region. Given the continuous nature of the outcome variable (rating) in the PAC, the selected models for exploration in this study included regression tree, bag_ipred, bag, forest, and forest_ranger. The choice of tree-based models was meticulously tailored to explore a spectrum of methodologies, each selected for its distinct characteristics and suitability to the dataset:

1. Regression Tree: Chosen for its straightforward approach and interpretability, the regression tree model is adept at decomposing complex decision-making processes into more manageable sub-decisions, thereby elucidating the path to the final prediction.

2. Bag_ipred: This model employs the bagging technique, particularly through the 'ipred' method, to enhance the stability and accuracy of predictions. Bagging is instrumental in mitigating issues of variance and overfitting, which are paramount in complex datasets.

3. Bag: Adopting a similar bagging approach but with a differing implementation from Bag_ipred, this model was selected to allow for comparative analysis within the bagging methodology, thereby enriching the study's exploration of ensemble techniques.

4. Forest: The inclusion of the random forest model was strategic, given its reputation as a robust ensemble method combining multiple decision trees to augment the overall predictive strength. Its efficacy in managing a multitude of input variables and resilience against overfitting made it an essential addition to the study.

5. Forest_ranger: A variation of the random forest model, Forest_ranger is renowned for its efficiency and scalability, particularly in handling extensive datasets. Its advanced features promised a potential enhancement in predictive accuracy, meriting its inclusion for evaluation.

The integration of these diverse tree-based models into the study was a deliberate strategy to encompass a range of analytical approaches. This selection was not only aimed at addressing the inherent non-linearity in the dataset but also at scrutinizing the efficacy of different tree-based methodologies in the realm of predictive analytics for music. By juxtaposing these models, the study sought to uncover their individual and collective strengths, as well as limitations, thereby contributing to a more nuanced understanding of predictive modeling in music analytics.

DOI: 10.54254/2755-2721/52/20241655

3. Data Analysis Process

3.1. Fail submission

During the experimental phase of this study, a meticulous process was undertaken to prepare the "analysisData" dataset for predictive modeling, encountering several challenges that provided key insights. Initially, data tidying was addressed, particularly in the 'genre' column, which was formatted in a disorganized manner. This involved the removal of extraneous characters like """, "[" and "]", utilizing the `stringr::str_remove_all` function for a cleaner and more analyzable dataset. Additionally, the 'id' column, deemed irrelevant for the prediction of song ratings, was removed to streamline the data.

The dataset also presented the challenge of missing data in the 'genre' column. Given the complexity of this variable—character type with a variable number of genres per observation—and the intricacies involved in imputing missing data, a decision was made to delete rows with missing genre observations. This approach, while simplifying the analysis, had unintended consequences in later stages.

In line with the insights gained from the EDA, various tree models were identified as suitable for the prediction task, with the initial focus on a random forest model named 'forest'. This model was applied to generate predictions, resulting in a new dataset of predicted song ratings. However, a significant obstacle emerged during the Kaggle submission process. The competition's submission guidelines stipulated a file containing exactly 4,844 rows with a header. The prior removal of rows with missing data led to a shortfall in the required row count, rendering the submission file incomplete and ultimately leading to the failure of the upload attempt.

This phase of the research highlighted the critical importance of comprehensive data preparation in predictive modeling. It underscored the delicate balance between effective data cleaning and the need to maintain dataset completeness, especially in the context of adhering to specific submission criteria in a competitive environment.

3.2. Coping with failure and optimizing the model

In response the challenges encountered during the initial submission phase of this study, a comprehensive revision of the methodology was undertaken to enhance the predictive model's performance. This revision was particularly focused on the 'genre' column in the dataset, which had presented significant challenges in the earlier stages of the research.

1. Creation of Dummy Variables:

The introduction of dummy variables to represent the genre data marked a pivotal methodological shift. This approach was necessitated by the recognition that simply deleting missing values was not viable, particularly given the importance of genre in predicting song ratings. The genre column in the dataset was rich with diversity, encompassing approximately 500 unique elements. Each song in the dataset was associated with a varying number of these genre elements, adding to the complexity of the analysis. To address this, each genre element was transformed into a dummy variable. For instance, 'rap' was designated as a dummy variable; in cases where a song's genre included rap, the variable was set to 1, and 0 otherwise. This binary representation allowed for a more structured and analytical approach to genre classification. Out of the extensive genre list, the 65 most frequent genres were identified and utilized as dummy variables, ensuring that the most significant genres in terms of frequency were included in the model.

2. Model Development and Evaluation:

With the dummy variables established, the study progressed to the exploration and evaluation of various advanced tree-based models. The selection of these models was guided by their potential to handle the complex and multi-dimensional nature of the data. The primary objective was to assess and compare the performance of these models, with a particular emphasis on test Root Mean Square Error (RMSE) as a key metric of accuracy and effectiveness. This phase involved rigorous testing and validation processes, wherein each model's predictive capacity was scrutinized and compared against its counterparts. The evaluation process was meticulous, ensuring that the findings were robust and reliable.

DOI: 10.54254/2755-2721/52/20241655

3. In-depth Analysis of Genre Impact:

The decision to focus on genre as a critical feature in the model was based on the hypothesis that different genres have varying levels of appeal and popularity among listeners. By converting genre elements into dummy variables, the study could delve deeper into understanding how specific genres influence song ratings. This analysis provided rich insights into the nuances of musical preferences, highlighting trends and patterns that were previously obscured.

4. Challenges in Model Optimization:

While the introduction of dummy variables for genre representation significantly improved the model's capability, it also introduced new challenges in terms of computational complexity and model interpretability. Balancing the accuracy of the model with its interpretability became a key focus, ensuring that the final model was not only effective but also comprehensible in its predictions. The comprehensive approach to data preparation and model optimization in this phase of the research was instrumental in overcoming initial setbacks. It demonstrated the importance of adaptive methodology in predictive analytics, especially when dealing with complex datasets like those in music. The nuanced inclusion of genre data, coupled with the rigorous evaluation of advanced tree models, laid a strong foundation for accurate and reliable song rating predictions, contributing significantly to the field of music analytics.

3.3. Best Prediction

In the pursuit of identifying the most effective predictive model for song ratings, a comparative analysis of test RMSE across five different models was conducted. This analysis revealed that the 'forest ranger' model yielded the lowest RMSE, indicating its superior predictive accuracy. Consequently, the 'forest ranger' model was identified as the optimal choice for this study, demonstrating its efficacy in accurately predicting song ratings based on the dataset's auditory features. This finding underscores the importance of rigorous model comparison and validation in predictive analytics, particularly in the context of complex datasets like those used in music analysis.

Table 1. RMSE for different Tree Models

Model	RMSE in Test
regression tree	15.46826
bag_ipred	15.44286
bag	14.7458
forest	14.72337
forest_ranger	14.67182

4. Discussion

In this research, while significant strides have been made in applying tree-based models for predicting song ratings, the study reveals several areas necessitating methodological improvements and refinement. A primary area for enhancement is the expansion of the range of predictive models. The reliance on a specific set of tree-based models, though effective, may not fully capture the nuanced relationships within music data. Future research could benefit from exploring a broader array of models, particularly advanced algorithms like deep learning, which have shown promise in handling complex, high-dimensional datasets.

The strategy for handling missing data in this study, primarily through deletion, is another area that requires a more nuanced approach. This method, while straightforward, may lead to potential biases or the loss of valuable information. Advanced techniques for missing data imputation, such as multiple imputation or model-based approaches, could provide a more accurate representation of the dataset and, consequently, more reliable predictive outcomes.

Additionally, the dynamic and subjective nature of music genres warrants a more nuanced treatment than the current method of dummy variable classification. Incorporating methods like natural language

DOI: 10.54254/2755-2721/52/20241655

processing could more accurately reflect the fluidity of genres. The temporal dynamics of music preferences, an element not extensively covered in this study, present another avenue for future research, potentially through time-series analysis.

These identified areas for improvement highlight the potential for methodological advancements in future research. By addressing these aspects, subsequent studies can build upon the current research's foundations, enhancing the accuracy and applicability of predictive models in music analytics.

5. Conclusion

This research project, centered on participating in the PAC, culminated in several key insights that are integral to the field of predictive analytics, particularly in music, which not only contributed to the understanding of predictive model selection in music analytics but also provided practical insights into data preprocessing and model evaluation techniques. These findings have broader implications for the field, offering guidance for future research and applications in predictive analytics within the music industry.

The first enlightenment pertains to the choice of predictive models; it was observed that when the distribution of independent and dependent variables exhibits a spherical shape in scatter plots, tree models generally outperform linear regression models in terms of predictive accuracy. This underscores the importance of selecting appropriate modeling techniques based on data distribution characteristics.

The second insight relates to data preprocessing, specifically the treatment of missing values. The study highlighted that the direct deletion of missing values is not always the most effective strategy, as it may lead to incomplete datasets and adversely affect the predictive model's performance. This emphasizes the need for careful consideration and innovative approaches in handling missing data.

Lastly, the research demonstrated the utility of comparing the predictive effectiveness of various models using test RMSE. Among the five tree models evaluated in this PAC, the 'forest_ranger' model emerged as the most effective, exhibiting the lowest test RMSE and thereby aligning with the project's goal of accurately predicting song ratings.

References

[1] C. Kern, T. Klausch, and F. Kreuter, "Tree-based Machine Learning Methods for Survey Research," Surv Res Methods, vol. 13, no. 1, pp. 73-93, Apr. 11, 2019, PMID: 32802211; PMCID: PMC7425836.

[2] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement," Information, vol. 11, no. 6, Art. no. 332, 2020. [Online]. Available: https://doi.org/10.3390/info11060332

[3] S. Dubnov, G. Assayag, O. Lartillot and G. Bejerano, "Using machine-learning methods for musical style modeling," in Computer, vol. 36, no. 10, pp. 73-80, Oct. 2003, doi: 10.1109/MC.2003.1236474.

[4] E. Nakamura, M. Hamanaka, K. Hirata and K. Yoshii, "Tree-structured probabilistic model of monophonic written music based on the generative theory of tonal music," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 276-280, doi: 10.1109/ICASSP.2016.7471680.

[5] K. West, S. Cox, and P. Lamere, "Incorporating machine-learning into music similarity estimation," in Proc. 1st ACM Workshop on Audio and Music Computing Multimedia (AMCMM '06), New York, NY, USA, 2006, pp. 89–96.