Melody Metrics: A Comparative Analysis of Predictive Models in Song Popularity

Anjaneya Ketkar
University of Maryland, College Park
College Park, USA
aketkar@umd.edu

***Abstract*—Predicting the popularity of a song upon release is a critical decision for music studios, guiding their investment in promotional efforts. This journal article conducts a comparative analysis of machine learning algorithms using the Spotify API dataset to enhance the accuracy of such predictions. Notably, the study reveals compelling results, with decision tree and linear discriminant analysis models exhibiting commendable accuracy. The decision tree model, selected based on Receiver Operating Characteristic curve and number of false positives, emerges as a particularly robust choice. Subsequently, the paper explores avenues for further enhancements in predictive capabilities, providing valuable insights for the music industry's strategic decision-making processes.**

***Index Terms*—Logistic Regression, Decision Tree, KNN, Naive Bayes, Linear Discriminant Analysis classifiers , Receiver operating characteristic Curve, Spotify API Dataset, K-Fold Cross Validation**

I. Introduction

Music studios face a challenging task of predicting the popularity of a song before its release. This prediction affects their strategic decisions on how to allocate their budget and time for marketing campaigns, as well as whether to sign the artist who produced the song. Missing a potential hit song can result in a significant loss of revenue from online platforms such as YouTube and Spotify. Therefore, there is a need for an effective and reliable method to identify and quantify the factors that influence the popularity of a song.

In this study, various machine learning algorithms like decision tree, logistic regression, K's Nearest Neighbors, Naive Bayes and Linear Discriminant Analysis algorithms are used to train on the dataset and make prediction on whether a song would be popular or not. Then in order to evaluate the model more precisely the confusion matrix and area under the Receiver Operating Characteristic curve is calculated to further check which algorithm works the best. The algorithm with the best results will be used by the music studio to predict the songs popularity.

A. Literature Review

Predicting the success of a song has garnered considerable attention within the research community. In the work presented by [1], apriori parameters are utilized for model training, incorporating data gathered from diverse sources. Despite the comprehensive approach employed, this study takes a distinct stance by asserting that the complexity of predicting a song's success cannot be fully addressed through the lens of data science alone.

Contrastingly, in the domain of Dance Hit Song Prediction [2], researchers trained their model spanning the years 1983 to 2013, successfully predicting hits in 2015 and validating these predictions against Billboard rankings. Remarkably, this study achieved a high level of accuracy in its predictions.

B. Dataset

The Dataset used in this paper is present on kaggle as Spotify Dataset 1921-2020[3]. This consists of separate files for a ten year set. In this paper dataset from 2000 to 2020, consisting for 12269 records, is used. The dataset consists of various columns like track name, artist name, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, ,time_signature, chorus_hit, sections and target. For a more detailed overview, additional information can be referenced on the Spotify API developer website [4].

This paper is organized into five key sections to provide the approach to predict the whether a song would be popular or not. The 'Introduction' sets the stage by introducing the problem, the importance of solving it, literature review and details about the dataset. 'Methodology' delves into the technical implementation details, covering model initialization, and training the model and how the modes were validated. In 'Results,' the authors present the empirical outcomes of the models on the dataset and comparison between different models. 'Discussion' critically examines these results and provides a comparative analysis of these models with logistic regression, decision tree, KNN algorithms, naive bayes and Linear Discriminant Analysis models and suggests potential directions for future research.

II. Methodology

The logistic regression, decision tree, KNN, Naive Bayes and Linear Discriminant algorithms were trained on 80% the spotify dataset and then tested on the remaining 20%. Then cross validation was applied to check the robustness of the model. Further, Area under the Receiver operating characteristic Curve was calculated to choose the best model for the spotify dataset.

A. Versions Used

The methodology employed in this study is grounded in Python, utilizing version 3.11.2, and relies on crucial libraries such as NumPy (version 1.25.2), sklearn (version 1.3.0), and Pandas (version 2.1.0) for data manipulation and analysis.

B. Processing Data

The data from the csv was converted into pandas dataframe using pd.read_csv() function present in the pandas library[5]. This was loaded into two separate frames and then they were concatenated using the concat() method[6].

Listing 1. Preprocessing Data

```
datasetFor2000 = pd.read_csv(
                    "dataset-of-00s.csv")
datasetFor2010= pd.read_csv(
                     "dataset-of-10s.csv")
x=pd.concat([datasetFor2000,
                datasetFor2010])
```

The pd.read_csv() function takes the csv file as it's parameter and uses ',' as it's default separator[5]. The concat function takes two or more pandas dataframe as their required parameter and many optional parameters[6]. Then few of the parameters like the song name, song artist and song uri are dropped using drop function in pandas library as done in[11].

In this study, the machine learning models employed—decision tree, logistic regression, K's nearest neighbours, Naive Bayes and Linear Discriminant Analysis models upon descriptions, equations and methodologies previously elucidated in the earlier publication [7];[8];[9];[10]. For comprehensive insights into these models and their implementations, readers are encouraged to consult the referenced work for detailed descriptions and evaluations. For the KNN algorithm k values from 1 to 300 were checked to find the best value of K as done in [9].

Once, the accuracy's(percentage of correct predictions) of the models was calculated. The two best models were chosen and then cross validation was applied to these models to check their robustness. This cross validation definition and usage is similar as[12].

Further, in order to get a better idea about the true positive and false positives the confusion matrix and Area Under the Receiver operating characteristic Curve was calculated. Detailed information on the description, initialization, usage, and plot of this metric can be found in [11] and [13].

III. Results

The decision tree yielded an accuracy of 77%, the logistic regression achieved 51%, KNN reached 67%, Naive Bayes obtained 65%, and the Linear Discriminant Analysis model demonstrated the highest accuracy at 79%.

For KNN, the optimal K value was determined to be 170. This choice is evident in Figure 1, where the plot exhibits a noticeable flattening around this value. The linear discriminant analysis, and decision tree have performed better than the rest and thus require further analysis. The five fold cross validation result for linear discriminant analysis was 0.78443358 0.80399348 0.799511 0.79502852 0.79331431 and decision tree was 0.77302363 0.77098615 0.76731866 0.76609617 0.78108439. The values for all the models are very close to each other and hence the models are robust and not over-fitted.

TABLE I
Accuracy Scores of the Five Models

Model ↓	Accuracy
Decision Tree	77%
Logistic Regression	51%
KNN	67%
Naive Bayes	65%
Linear Discriminant Analysis	79%

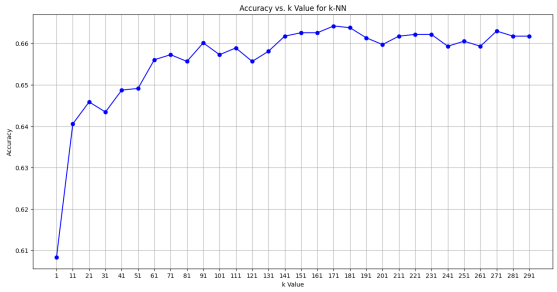

Fig. 1. KNN for values of K from 1 to 300

The confusion matrix analysis for Linear Discriminant Analysis (LDA) revealed that it correctly predicted 841 hits and accurately identified 1141 records as not being hits. However, it missed out on 417 hits and misclassified 97 records as hits when they were not while for the Decision Tree model correctly predicted 950 instances as hits and 948 instances as non-hits. However, it misclassified 284 hits as non-hits and 272 non-hits as hits.

The area under the curve for the decision tree was found to be 0.77 while for the linear discriminant it was 0.80. However, the slope of the curve for decision tree was much more steeper than the slope for the Linear discriminant analalysis as it can be seen in figure 2.

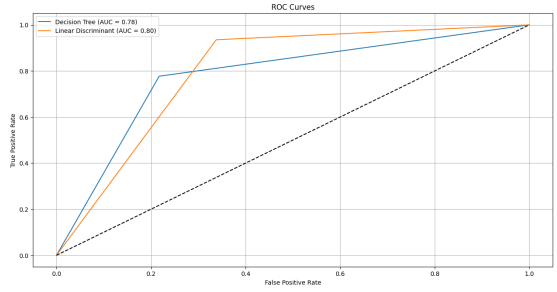

Fig. 2. AUC ROC curve for LDA AND DT

IV. Discussion

The evaluation of various machine learning models has revealed that both the Linear Discriminant Analysis (LDA) model and the Decision Tree model outperform the other three models in predicting hit songs, yielding comparable accuracies of approximately 79%. Upon conducting a detailed analysis of cross-validation results, it is evident that both models exhibit robust performance across diverse subsets of the data. This consistency underscores their generalizability and reliability in predicting hit songs. Now faced with the decision to prioritize either reducing false positives or false negatives, the music company must consider the substantial costs associated with music procurement, production, and marketing. Given the high stakes involved, careful consideration is crucial in selecting the model that aligns with the company's strategic objectives and risk tolerance.

The dilemma intensifies as the production companies navigate the trade-off between minimizing the risk of investing in non-hit songs and the potential rewards of a hit. The financial implications of missing out on a hit song are emphasized by the staggering opportunity cost, where the rewards of a single hit could outweigh the investment in producing a hundred non-hit songs and also the company might miss out on an album of songs where the LDA model would falsely-predict many non-hit songs. Delving into the specifics of the confusion matrix, it becomes apparent that the LDA algorithm excels in identifying hit songs but at the expense of missing 30-35% of potential hits. Recognizing this drawback, the production company may lean towards the Decision Tree model, which, while maintaining competitive accuracy, offers a more balanced trade-off between identifying hits and avoiding false negatives. In conclusion, the decision to opt for the Decision Tree model aligns with the company's imperative to strike a prudent balance between cost considerations and the potential windfall associated with producing a hit song.

This paper contributes significantly to the evolving landscape of predicting hit songs, providing valuable insights for music companies in their decision-making processes for song promotion. By leveraging machine learning models, the research aids in identifying potential hit songs, offering a strategic advantage to music companies aiming to maximize revenue. The findings emphasize the importance of a careful balance between minimizing false negatives and the substantial rewards associated with producing hit songs. The recommendations provided can guide music companies in optimizing their song selection processes, ultimately enhancing their ability to identify and promote songs with the potential for significant commercial success.

Looking ahead, the future of hit song prediction could benefit from the incorporation of more advanced and intricate models. To improve accuracy and refine metrics, exploring sophisticated machine learning algorithms and techniques could be instrumental. Additionally, enhancing the dataset by incorporating additional fields, such as lyrics analysis, could provide a more comprehensive understanding of the factors influencing a song's success. Assigning points based on the popularity potential of keywords within lyrics and considering the song's suitability for trending on platforms like TikTok and Instagram Reels could further enhance the predictive capabilities of the model. By embracing a more multifaceted approach, future research in this field has the potential to offer even more nuanced and accurate predictions, providing greater value to the music industry in its quest to identify and promote hit songs effectively.

References

[1] Krishnadas Nanath and Agha Haider Raza :Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?

[2] Dorien Herremans:Dance Hit Song Prediction

[3] https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks

[4] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

[5] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

[6] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.concat.html

[7] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, Classification using Logistic Regression

[8] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, Decision Tree Analysis on Iris Dataset

[9] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, K-Nearest Neighbor: A Comprehensive Study on 'K' and Decision Metrics

[10] A. Ketkar: Classification of Iris dataset using Naive Bayes and Linear Discriminant analysis classifiers

[11] A. Ketkar:Selecting Flight Delay Prediction Models Using ROC AUC Evaluation

[12] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu,Comparing Decision Tree with Logistic Regression using Cross-Validation Technique

[13] A. Ketkar:Maximizing Profit in Iris Flower Sales: A Comparative Analysis of Machine Learning Algorithms