Harmonizing Hits: Forecasting Song Popularity

Chandranshu Verma
cverm078@uottawa.ca
300372673

Abstract—Predicting song popularity is critical in the dynamic music industry, necessitating a thorough understanding of influencing factors. Leveraging the Song Popularity Dataset, encompassing audio features and metadata for around thirteen thousand songs, I employed various classification algorithms to assess their efficacy in predicting popularity. I analysis focused on identifying the most influential features within the dataset. The study provides valuable insights for music industry professionals, shedding light on predictive algorithms and key factors driving a song's success.

I. INTRODUCTION

This project delves into predicting song popularity by leveraging machine learning algorithms and analyzing diverse characteristics of songs. The significance of understanding the determinants of song popularity is emphasized, considering the substantial impact of the music industry on human culture and the multi-billion-dollar revenue it generates. This research aims to contribute to the evolving field of "Hit Song Science," which utilizes machine learning to predict song popularity. By employing various algorithms, including SVC, logistic regression, Random Forest, and Gradient Boosting Classifier, I seek to accurately determine whether a song will achieve popularity without using the score of song popularity. The outcomes of this research are expected to provide valuable insights for music-related businesses, such as radio stations, record labels, and digital/physical music marketplaces, while also contributing to the enhancement of personalized music recommendations.

II. RELATED WORK

The significance of social impact in influencing a song's popularity has been thoroughly investigated in research on popularity prediction. Bertin-Mahieux et al. showed how to use AdaBoost and FilterBoost in machine learning approaches to label music based on acoustic data [1]. SVMs, as an alternative to AdaBoost with FilterBoost, may have been a better choice, nonetheless.

Using a variety of regression and classification techniques, Koenignstein, Shavitt, and Zilberman estimated billboard success based on peer-to-peer networks and captured the social influence on song popularity [2].

A more upbeat take on predicting music popularity was provided by Ni et al. [3], who classified the top 5 singles from the top 30–40 hits using a Shifting Perceptron algorithm. Their study yielded better outcomes since it contained more innovative audio characteristics.

Pachet and Roy contested the notion that cutting-edge machine learning cannot be used to understand song popularity [4]. Their study took into account metadata as well as auditory information, however it also included a large amount of features without stating a feature selection algorithm, which could cause overfitting.

Salganik, Dodds, and Watts discovered that social influence had a significant impact on a song's popularity, with quality having a minor effect [5]. The objective of this project is to enhance the model's prediction accuracy by integrating both metadata and audio elements.

III. DATASET AND FEATURES

A. Data

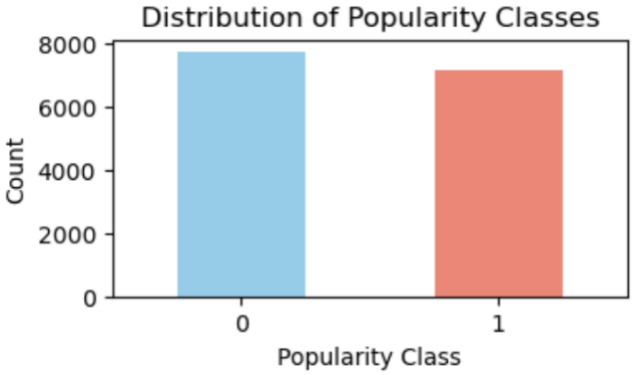
This study utilized the Song Popularity Dataset [6], an extensive compilation of audio attributes and accompanying metadata for nearly thirteen thousand songs. This dataset encompasses a comprehensive set of attributes, including details about the music track such as duration, key, and name, as well as more abstract features like danceability, energy, and loudness, popularity. Below are the column headers of the data:

song_name, song_popularity, song_duration_ms, acousticness, danceability, energy, instrumentalness, key, liveness, loudness, audio_mode, speechiness,tempo, time_signature, audio_valence

To ensure the robustness of our analysis, we divided the tracks into two sets: 75% of the tracks were allocated for training, and the remaining 25% were reserved for testing. This division allows us to train our predictive models on a substantial portion of the data and evaluate their performance on a separate, unseen dataset, providing a reliable assessment of the models' predictive capabilities.

B. Features and Target Variable

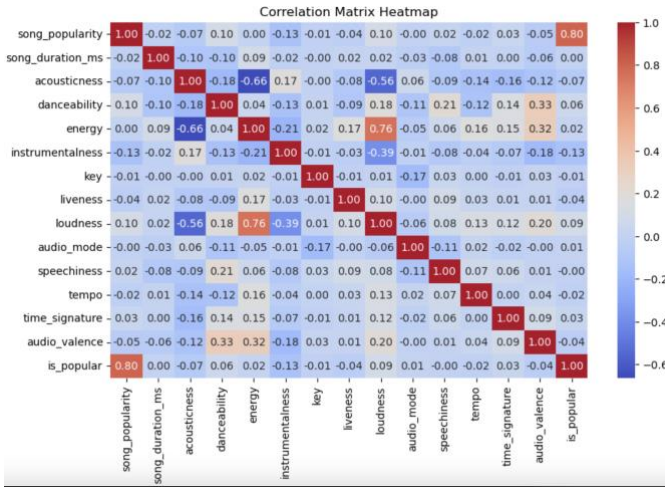
1) Target Variable: In this dataset, I introduced a new column called "is_popular," derived from the "song_popularity" feature. The classification is based on whether a song's popularity surpasses the mean popularity of the entire dataset. If a song's popularity is greater than the mean, it is labeled as "1"; otherwise, it is labeled as “0." The primary objective of our analysis is to predict and understand the factors influencing whether a song falls into the category of being popular or not.



2) Features: The dataset encompasses a diverse set of features, including song name, popularity, duration, acousticness, danceability, energy, instrumentalness, key, liveness, loudness, audio mode, speechiness, tempo, time signature, and audio valence.

IV. METHODS

A. Feature Selection



In this music popularity prediction project, delving into the intricacies of feature relationships was crucial for effective feature selection during model training, and we utilized a correlation matrix for this purpose.

Positive correlation was observed, indicating that an increase in one feature corresponds with a rise in another. For example, a positive correlation between "loudness" and "is_popular" suggests that more louder songs tend to be more popular. Perfect positive correlation (coefficient of 1) means a 100% increase in one feature corresponds to an increase in the other.

Conversely, negative correlation was explored, revealing that an increase in one feature leads to a reduction in another. A negative correlation between "instrumentalness" and "is_popular" suggests less acoustic songs might be more popular. A correlation coefficient of -1 indicates a perfect negative correlation.

In predicting song popularity, features like "Danceability," "Energy," and "Loudness" are significant. Caution against over-selecting features aims to prevent increased training time and model complexity, guarding against overfitting. The goal is to identify the optimal feature set that balances model complexity with predictive performance, enhancing accuracy on new, unseen data in our music popularity prediction model.

Also after training and testing the model, I used feature importance to see with which all features affect the most.

After seeing both correlation matrix and feature importance, the features finalised are:
song_duration_ms, acousticness, danceability, energy, instrumentalness, liveness, loudness, audio_valence

B. Classification

In quest to predict music popularity, it is better to classify the song as popular or not rather than predicting the score of poularity as it is easy to understand and the selection of a suitable machine learning model is also a pivotal decision. For the task of discerning whether a song is popular or not, I opted for a variety of classifiers tailored to the nuances of objective.

I find the **Random Forest Classifier** to be particularly versatile, excelling in handling non-linear relationships and intricate feature interactions. Leveraging an ensemble of decision trees, this model ensures robustness and high predictive performance, making it well-suited for scenarios where the relationship between features and song popularity involves complex, non-linear patterns.

Opting for **Logistic Regression**, a simpler yet interpretable model, serves as an ideal starting point for binary classification tasks like ours. Particularly effective when the relationship between features and popularity is approximately linear, Logistic Regression provides valuable insights into the impact of individual features on predictions. Its simplicity enhances interpretability, offering a straightforward understanding of feature influence.

The **Support Vector Classifier (SVC)** proves to be a formidable choice for classification tasks, especially when dealing with non-linear relationships. Robust in high-dimensional spaces, SVM excels at capturing intricate relationships within the data. Its effectiveness becomes particularly valuable in our music popularity prediction task, where the relationship between features and popularity may involve non-linear patterns.

I also considered the **Gradient Boosting Classifier**, represented here by XGBoost, as a powerful alternative known for outperforming other models. Ideal for capturing non-linearity and feature interactions, Gradient Boosting provides insights into feature importance, crucial for understanding the nuances of song popularity prediction. While demanding more hyperparameter tuning, the enhanced predictive performance justifies the effort, making XGBoost particularly effective in our context.

V. EXPERIMENTAL RESULTS

In exploration of predicting music popularity, I conducted a comprehensive analysis, experimenting with diverse model configurations. The Random Forest Classifier, employing 100 trees with a maximum depth of 5, achieved 59.27% accuracy. It demonstrated balanced precision, recall, and F1-

score, indicating effective pattern capturing without overfitting.

The Logistic Regression model, with 56.30% accuracy, showed better performance for popular songs (class 1). Its interpretability is valuable, but the moderate accuracy suggests potential for improvement.

The Support Vector Classifier (SVC) excelled with 58.74% accuracy, demonstrating a balanced predictive ability for both classes. Leveraging an RBF kernel, the SVC effectively captured non-linear relationships, with good generalization.

The Gradient Boosting Classifier, with 58.63% accuracy, showcased balanced performance for both classes, handling non-linear relationships effectively. The reasonably close training and test accuracies indicated good generalization.

Collectively, these classifiers provide diverse approaches to predicting music popularity, each with unique strengths.

VI. CONCLUSION

This project revolved around predicting music popularity without using the score of song popularity, utilizing diverse machine learning classifiers such as Random Forest, Logistic Regression, Support Vector Classifier (SVC), and Gradient Boosting Classifier. Following meticulous configuration and evaluation, the classifiers demonstrated varying performances. Notably, the Random Forest classifier emerged as the top performer after hyper tuning it various times, achieving a notable accuracy of 59.27%. This signifies that Random Forest excels in capturing complex patterns in the data, outperforming other classifiers.

While the Logistic Regression model showcased an accuracy of 56.30%, particularly excelling in predicting popular songs (class 1), and the Gradient Boosting Classifier achieved an accuracy of 58.63% with balanced precision, recall, and F1-score metrics, the dominance of Random Forest in terms of model accuracy establishes it as the classifier of choice for our music popularity prediction task. Further optimization and fine-tuning of the Random Forest model may enhance its already impressive performance. These results collectively underscore the potency of machine learning in uncovering intricate patterns in music data and predicting popularity, showcasing the applicability of these models to real-world scenarios through their ability to generalize well to new, unseen data.

In summary, this project provided a thorough evaluation of different classifiers, ultimately identifying Random Forest as the most accurate model. These findings significantly contribute to

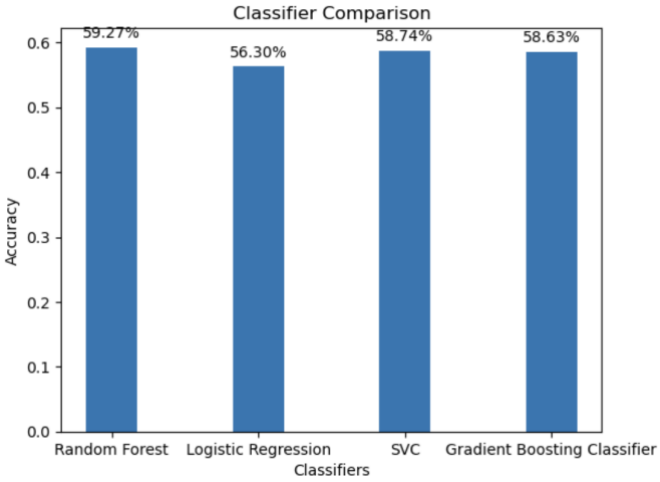
the field of music recommendation systems, offering valuable insights for stakeholders in the music industry and advancing the understanding of the multifaceted factors influencing song popularity.

```
Random Forest Accuracy: 59.27

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.60      0.61      1945
           1       0.57      0.58      0.58      1787

    accuracy                           0.59      3732
   macro avg       0.59      0.59      0.59      3732
weighted avg       0.59      0.59      0.59      3732
```



REFERENCES

[1] Bertin-Mahieux, Thierry, et al. ”Autotagger: A model for predicting social tags from acoustic features on large music databases.” Journal of New Music Research 37.2 (2008): 115-135.

[2] Koenigstein, Noam, Yuval Shavitt, and Noa Zilberman. ”Predicting billboard success using data-mining in p2p networks.” Multimedia, 2009. ISM’09. 11th IEEE International Symposium on. IEEE, 2009.

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

[4] Pachet, Franois, and Pierre Roy. ”Hit Song Science Is Not Yet a Science.” ISMIR. 2008.

[5] Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. ”Experimental study of inequality and unpredictability in an artificial cultural market.” science 311.5762 (2006): 854-856.

[6] Yasser H. "Song Popularity Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/yasserh/song-popularity-dataset.

[7] J. Pham, E. Kyauk, and E. Park, "Predicting Song Popularity," Stanford University, Department of Computer Science, Tech. Rep., 2015. [Online]. Available: https://cs229.stanford.edu/proj2015/140_report.pdf