SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Predicting Hit Songs:
An Exploratory and Comparative Study

By Yanrui Jerry Wu

AUTHOR BIOGRAPHY

Jerry Wu is currently a senior at Basis International School Park Lane Harbour. He is particularly interested in data science, finance, and music. Participation in the Wharton Global Youth Investment Challenge highlighted how data science informs decisionmaking in finance, further piquing his interest in the field of data science and machine learning. Additionally, he is also the lead singer of his original rock band – Six Drowners, which has over 15 original songs written and performed – and the co-founder of YouMe, a non-profit organization dedicated to organizing high school student live house events in Shenzhen. With these experiences across different academic fields, Jerry Wu aims to research interdisciplinary projects, further exploring his interest in data science and music.

ABSTRACT

Data science is currently at the forefront of the music industry, especially in streaming media companies like Spotify as songs' audio features can be labeled with specific values and weights. Using the Kaggle Spotify Tracks Dataset, a labeled dataset in which hits are defined as tracks at or above the 80th-percentile popularity is assembled to train the models. A comprehensive exploratory data analysis characterizes the nonlinear distributions and indicates the correlations between audio features and song popularity. Four classifiers—Logistic Regression, Support Vector Machine, Random Forest, and Neural Network—are trained in standardized pipelines. To align with the discovery objective of optimizing the genuine "hit" prediction, the study applied class-weight balancing and cost-sensitive threshold tuning to maximize recall. The results demonstrate that a tree-based ensemble, particularly Random Forest, gives the best precision-recall balance, while SVM also outputs a competitive recall under a tuned threshold. The study explores the connection between audio features from an objective statistical standpoint and identifies that tree-based ensembles perform better when dealing with tabular datasets.

Keywords: *Data Science, Music, Exploratory Data Analysis, Machine Learning, Trend Analysis, Statistics, Kaggle, Spotify*

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

INTRODUCTION

Predicting whether a song will become a "hit" is valuable for artists, labels, and streaming platforms. Prior studies suggest that a song's audio characteristics can predict its mainstream appeal (Lee & Lee, 2018; Interiano et al., 2018). Building on this, our study addresses a specific question: to what extent do the readily available audio features alone predict a hit/non-hit trend of a song? Our objective is twofold: to train supervised models for hit-song prediction and to demonstrate a complete data science workflow, from exploratory data analysis to model evaluation.

To accomplish the goal, a dataset of songs from Kaggle's *Spotify Tracks Dataset* (Pandya, n.d.) is assembled to analyze audio features (e.g., danceability, energy, loudness, valence) and ask whether these characteristics can predict a binary "hit" label, which is defined as the top 20% popularity in the dataset, tracks at or above the 80th percentile are labeled 1 (hit), others 0 (non-hit). Then, four supervised classifiers—Logistic Regression, Support Vector Machine, Random Forest, and Neural Network— are compared under standardized evaluation.

DATA AND METHODS

Data

The dataset is assembled via the Spotify Song API, randomly collecting approximately 114,000 songs. After deduplicating the dataset, each specific track contains 20 features categorized by track information, artist information, album information, and audio analysis features. The predicting label, hit song, is defined by the top 20% popularity in the dataset. This dataset is suitable because it provides enough audio features for analysis with a given label for testing. Additionally, containing songs from diverse genres (125 different genres) with different audio features and having the song distribution across different genres being roughly uniform, the dataset ensures its randomness and representativeness. Therefore, the dataset satisfies with the interest in music analysis and provides a fairly representative dataset to train the models to predict song popularity.

I describe these features in detail below:

track_id: the song's unique Spotify track ID.
artists: the artist of the song.
album_name: the name of the album that the song is in.
track_name: the actual name of the song.

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

popularity: the popularity score from 1 to 100 calculated by Spotify, with 0 being the least popular and 100 being the most popular, and is calculated by "the total number of plays the track has had and how recent those plays are (usually a 30-day period)." (Spotify for Developers, n.d.)
duration_ms: the duration of the track in milliseconds.
explicit: whether the track is explicit or not, with 0 being non-explicit, and 1 being explicit.
danceability: numerical value (ranging from 0.0 to 1.0) that indicates how suitable a song is for dancing, based on elements including tempo, rhythm stability, beat strength, and overall regularity. A higher danceability score suggests the track is more likely to get people to dance.
energy: a value from 0.0 to 1.0 that represents a perceptual measure of intensity and activity.
key: the key the track is in.
loudness: a numerical value measuring the loudness of the track in decibels (dB).
mode: the modality (1 = major or 0 = minor) of a track
vocal content: a numerical value from 0.0 to 1.0 that indicates the overall words spoken in the track. A value of 0.0 implies that the song has no words included, for instance, a purely musical piece.
acoustic quality: a numerical value from 0.0 to 1.0 measuring whether the song is acoustic.
instrumental focus: a numerical value from 0.0 to 1.0 describing the use of instruments in the track. A higher instrumental score suggests that the track utilizes lots of instruments, for example, an orchestra.
liveness: a numerical value from 0.0 to 1.0 measuring the presence of the audience in the track. A higher liveness suggests the track is recorded during the performance with lots of audience around.
valence: a metric that represents the musical positiveness of a track, ranging from 0.0 to 1.0. Higher valence scores indicate more positive, "happy" sounding music.
tempo: a numerical value measuring the overall tempo of the track in beats per minute (bpm).
time_signature: an estimated signature of duration of the track.
track_genre: the classified genre of the track.

To build the dataset, I first resolved track IDs via search/playlist endpoints, then fetched audio features with the audio-features endpoint and metadata (including popularity). I kept only rows with complete features, deduplicated exact track IDs, used the "global" market when present, and removed duplicate releases of the same track. As mentioned above, Spotify's popularity index is an algorithmic, time-varying measure largely based on the total plays in recent time, so the label has some built-in noise.

EDA Highlights

Before training models, a comprehensive exploratory data analysis of the dataset is performed to better understand the relationships between audio features and song popularity and the correlation between each audio feature. The audio feature analysis indicates that hits and non-hits exhibit statistically significant differences across multiple musical characteristics, with hits displaying a high degree of danceability, energy, and

SCHOLARLY REVIEW JOURNAL

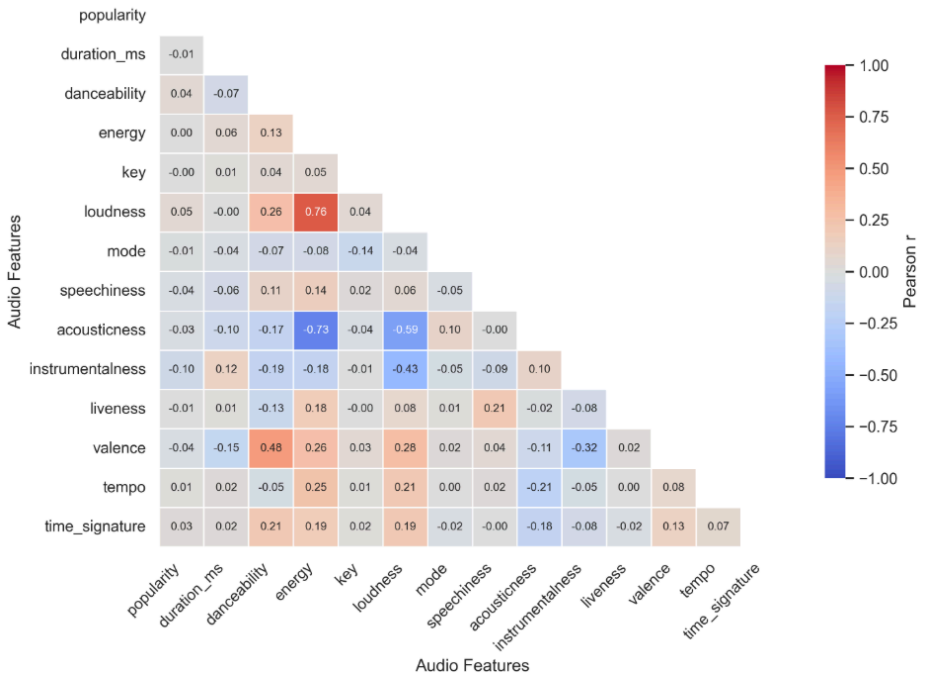
E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

loudness, while also having a lower level of instrumental focus and acoustic quality. Such a relationship is understandable as the correlation analysis reveals that the heat map shows that energy has a fairly strong positive correlation (0.76) with loudness, and a positive relationship exists between loudness and danceability (see Figure 1). The correlation analysis demonstrates a similar result: danceability, loudness, and tempo are the three most predictive features for song popularity, with loudness having the strongest positive correlation with popularity ($r \approx 0.5$), suggesting louder songs usually have higher popularity (see Figure 2). Genre performance analysis reveals significant variation in hit rates across different musical genres, with certain genres like pop-film and pop showing significantly higher success rates than others (see Figure 3), indicating the potential popularity trend in today's music industry. These findings provide an empirical foundation for machine learning models to predict the actual "hit" based on objective audio features instead of subjective popular artists.

Figure 1

Audio features correlation heat map, demonstrating the relationship between each audio feature.



SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Figure 2

The correlation of audio feature with song's popularity, with positive indicating directly proportional relationship between audio feature and popularity.

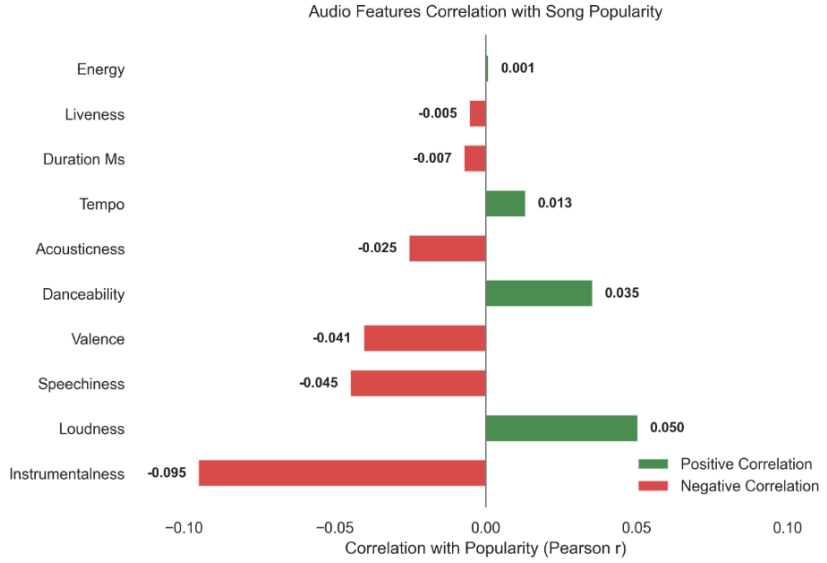
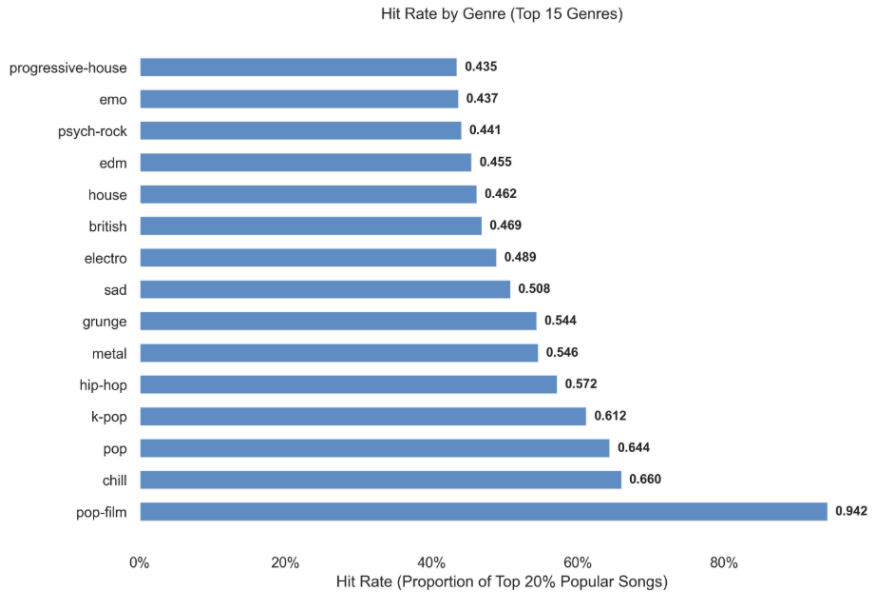


Figure 3

The top 15 genres ranking by popularity rate.



SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Models Considered

Four different classifiers are selected to span a range of interpretability and capacity for nonlinear relationships in numeric audio features:

Logistic Regression (LR): It describes a regularized linear baseline with interpretable coefficients that indicate the direction and relative strength of each feature's association with the "hit" label.

Support Vector Machine (SVM): A margin-based non-linear model that captures curved decision boundaries without explicit feature engineering, which might be useful in the scenario where audio features of the track are not linearly separable and related to some extent.

Random Forest (RF): An ensemble of decision trees that is useful for nonlinearities and interactions. This model is specifically useful in the case of hit prediction because it offers interpretability for the interaction between audio features and the hit label.

Neural Network (NN): A flexible model that captures complex and nonlinear relationships among audio features.

Rationale of the mix: These models collectively combine into a diverse system of models, covering different advantages and weaknesses. LR provides insights from a linear baseline with interpretable results; SVM and NN offer results from a flexible nonlinear point of view; and RF represents the decision tree focusing on tabular data. Coming together, these models offer a valuable evaluation in determining the best way to predict song popularity with its audio features.

Model Evaluation

I evaluated models with stratified 5-fold cross-validation, reporting mean ± standard deviation across folds. Initially, accuracy is the primary metric for my experimentation, and the results show very high accuracy with absurdly low precision and recall rates (below 30%). Reviewing the evaluation process, I found that since "hit" is the minority case, composing only 20% of the dataset. Therefore, valuing accuracy without proper sampling and balancing the dataset will result in deceptively high accuracy with low precision and recall due to predicting the majority class (successfully predicting the non-hit), which is not the main focus for the study. Accordingly, I prioritized the recall and F1 Score over the accuracy.

To address the data imbalance issue, three strategies are implemented. First, class weight balancing tells the model that the rare "hit" in the dataset weighs more than the common "non-hit" in the dataset, so each hit example counts extra during training. Second, under sampling temporarily removes some non-hit examples inside the training folds so the model sees a more even mix of both tracks. Third, a cost-sensitive setting penalizes

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

misclassification made by the model. The decision threshold is also tuned on each fold to maximize recall subject to a modest precision floor for the models that output probability. This emphasis on recall ensures the models are optimized in identifying the genuine "hits" in the real world, while avoiding being misled by the high accuracy of predicting the non-hits, which is not as useful as missing fewer hit songs to invest in the music industry.

RESULTS

The below table illustrates the results of each model with standard deviation (smaller Std means greater consistency)

From the results shown in Table 1, RF achieved the strongest overall balance (highest F1 with low variance), while SVM demonstrated the best recall score apart from RF, which is useful in the scenario where the primary focus is to capture as many potential hits as possible. LR underfits the nonlinear structure of hit prediction, as suggested by EDA. The NN's recall and F1 scores contain the highest variation across different folds. This is because, under the imbalanced tabular dataset and the 5-fold cross validation, there might be relatively few positive ("hit") labels per fold, making the feed-forward NN to fit fold-specific noise rather than general patterns, even with early stopping and class weights (Buda et al., 2018).

These results are consistent with each model's inherent nature. RF achieved a high and steady F1 score because it picks up feature interactions and, by averaging many trees, keeps variance low; there were also studies indicating RF outperforms other models on tabular data (Grinsztajn et al., 2022; Ye et al., 2025). SVM captures more true hits with class-weights and a lower decision threshold, which increases its recall (Rezvani et al., 2024). Since LR is a linear model, it misses nonlinear patterns of hit prediction. NN might be too complex for a tabular dataset in this case, which results in overfitting where the model is memorizing rather than learning the pattern (McElfresh et al., 2023). This ranking of models is also consistent with prior Spotify Feature study: on a similarly constructed dataset with both Spotify and Billboard data, RF and SVM are typically outperforming the other two models with higher precision-recall balance and accuracy (Middlebrook & Sheik, 2019).

Table 1

Cross-validation

Models	**Accuracy Mean ± Std**	**Precision Mean ± Std**	**Recall Mean ± Std**	**F1 Score Mean ± Std**
Logistic Regression (LR)	0.580 ± 0.003	0.569 ± 0.002	0.655 ± 0.008	0.609 ± 0.005

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Support Vector Machine (SVM)	0.633 ± 0.006	0.607 ± 0.004	0.752 ± 0.011	0.672 ± 0.007
Random Forest (RF)	0.728 ± 0.006	0.717 ± 0.006	0.756 ± 0.011	0.736 ± 0.006
Neural Network (NN)	0.638 ± 0.005	0.637 ± 0.012	0.642 ± 0.050	0.638 ± 0.021

Note. Stratified 5-fold cross-validation (mean ± standard deviation) results (Decision thresholds were tuned on validation folds to favor recall).

CONCLUSION

Through exploratory data analysis of the audio features, an interesting and important popularity trend in the music industry is revealed: recent trends prefer simpler (not involving too many instruments), louder, and more danceable music. Such a trend in popular music is also demonstrated by previous works, showing a growing loudness trend alongside homogenization of the timbral palette, meaning that popular music has become louder and simpler in its sonic makeup (Mauch et al., 2015, Serrà et al., 2012). These interpretations are also understandable as Figure 3 above displays the top 15 popular music genres – the populist genres in the figure, pop film, pop, and K-pop, usually have fewer instruments and are more danceable for people, which is consistent with the audio features analysis.

The study shows that using audio features alone provides feasible hit song prediction. To be specific, after balancing the dataset and using a recall-oriented evaluation of the models, decision-tree ensembles, especially **Random Forest**, outperform the linear and kernel models and offer the best model to predict the potential hit song. SVM gives the best recall value apart from RF, and NN presents the highest variation in cross validation.

However, there is one important leakage resulting from the same artist and the temporal effects. In the dataset, multiple songs by the same artist can affect the information across different folds, and the popularity of every track may change over time. To avoid the leakage, more stratified cross validation can be applied that groups artists in different folds. Since popularity is a time-varying label, this is noise that should be modeled and addressed specifically in future work.

Through this study, multiple technical lessons are learned. The most important lesson is about data preprocessing, balancing the hit and non-hit distribution in the dataset, since metrics may be obscured when the dataset is unbalanced. Next, the understanding of the emphasis of metrics is critical, which helps the model fit

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

better in the real-life scenario. Finally, evaluating the models with cross-validation is also significant, as it provides a more robust estimate of the model. For personal takeaways, designing and iterating on this project deepens my understanding of end-to-end machine-learning practice, from data analysis to model implementation and evaluation. Most importantly, I learned to frame the evaluation of a model critically around objectivity (real-life importance), instead of simply looking at the basic metric (accuracy in this case). Looking ahead, I will apply gradient boosted trees after seeing the performance of tree ensembles in this case, grouping by artist and temporal validation to achieve the precision/recall balance.

ACKNOWLEDGEMENTS

When writing this work, the author used ChatGPT.ai to refine and polish the sentence structure and grammatical errors in certain paragraphs. After using this tool, the author reviewed and edited the contents carefully and takes full responsibility for the publication of the essay.

REFERENCES

Lee, J., & Lee, J.-S. (2018). Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia, 20*(11), 3173–3182. https://doi.org/10.1109/TMM.2018.2820903

Interiano, M., Kazemi, K., Komarova, N. L., Wang, L., Yang & J., Yu, Z. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society Open Science, 5*(5), 171274. https://doi.org/10.1098/rsos.171274

Pandya, M. (n.d.). *Spotify tracks dataset* [Data set]. Kaggle. Retrieved August 8, 2025, from https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

Spotify for Developers. (n.d.). *Get artist's top tracks*. Retrieved August 3, 2025, from https://developer.spotify.com/documentation/web-api/reference/get-an-artists-top-tracks

Buda, M., Maki, A., & Mazurowski, M. A. (2018). *A systematic study of the class imbalance problem in convolutional neural networks.* Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *NeurIPS Datasets and Benchmarks Track*. https://arxiv.org/abs/2207.08815

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Ye, H.-J., Liu, S.-Y., Cai, H.-R., Zhou, Q.-L., & Zhan, D.-C. (2025). 300-dataset study shows method rankings cluster, with tree methods consistently strong on heterogeneous tabular data; DL success is dataset-dependent. *arXiv preprint.* https://arxiv.org/abs/2407.00956

Rezvani, S., Pourpanah, F., Lim, C. P., & Wu, Q. M. J. (2024). Methods for class-imbalanced learning with support vector machines: A review and an empirical evaluation. *arXiv preprint.* https://arxiv.org/abs/2406.03398

McElfresh, D., Guzmán-Rivera, A., Kim, S., Sculley, D., & Snoek, J. (2023). When do neural nets beat gradient boosting on tabular data? *NeurIPS (Datasets & Benchmarks).* https://arxiv.org/abs/2305.02997

Middlebrook, K., & Sheik, K. (2019). *Song hit prediction: Predicting Billboard hits using Spotify data*. arXiv. https://arxiv.org/abs/1908.08609

Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science, 2*(5), 150081. https://doi.org/10.1098/rsos.150081

Serrà, J., Corral, Á., Boguñá, M., Haro, M., & Arcos, J. L. (2012). Measuring the evolution of contemporary western popular music. *Scientific Reports, 2*, 521. https://doi.org/10.1038/srep00521