

Temporal Evolution of Hit Song Predictors from Spotify Audio Features: A Machine Learning and Explainable AI Study

Christopher Andreas^{1,*}, Felicia Sword²

^{1,2}*School of Information Technology, Universitas Ciputra, CBD Boulevard, Citraland Surabaya, 60219, Indonesia*

(Received: March 18, 2026; Revised: May 20, 2026; Accepted: September 5, 2026; Available online: October 14, 2026)

Abstract

Understanding the factors that distinguish commercially successful songs remains an important problem in music analytics, with implications for applied data science. While previous studies have demonstrated that machine learning models can predict hit songs from audio features, they generally treat the prediction task as static and pay limited attention to how the determinants of success evolve over time. This study investigates the temporal evolution of hit-song predictors using Spotify audio features extracted from 41,103 Billboard-labelled tracks released between 1960 and 2019. The dataset is divided into three industry eras (Analog, CD/Digital, and Streaming) to examine changes in predictive performance and feature importance across distinct periods of music production and distribution. Logistic Regression, Random Forest, and K-Nearest Neighbours are optimized using grid-search cross-validation and evaluated with accuracy, F1-score, and ROC-AUC. Model interpretability is achieved through SHAP analysis. Cross-era differences in feature importance are statistically evaluated using bootstrap confidence intervals, while robustness is further assessed through repeated random seeds, permutation importance, alternative era definitions, and data de-duplication. Among the evaluated models, Random Forest consistently achieves the highest predictive performance, with ROC-AUC ranging from 0.848 to 0.917 across the three eras, indicating that chart success has become increasingly predictable from audio characteristics over time. Explainable AI analysis reveals a systematic evolution in the importance of musical attributes: instrumentality remains the strongest predictor throughout, loudness becomes substantially more influential in the Streaming era, whereas speechiness and acousticalness progressively decline in importance. These patterns are supported by bootstrap statistical testing and remain stable across multiple robustness analyses, suggesting that the observed temporal changes are not artifacts of sampling variability, model specification, or data partitioning. Overall, this study demonstrates that explainable machine learning can characterize the temporal evolution of feature importance, providing a robust analytical framework for investigating long-term changes in cultural data.

Keywords: Feature importance, Hit song prediction, Music analytics, Random Forest, SHAP, Temporal evolution.

1. Introduction

The rapid expansion of digital music platforms has fundamentally transformed how music is produced, distributed, and consumed. According to the International Federation of the Phonographic Industry (IFPI), global recorded-music revenue reached US\$29.6 billion in 2024, representing a 4.8% annual increase, while paid streaming subscriptions exceeded 752 million users worldwide [1]. Despite this unprecedented growth, commercial success remains highly concentrated, with only a small proportion of released songs achieving sustained chart performance and attracting the majority of audience attention. Consequently, identifying the factors that distinguish commercially successful songs has become an important research challenge in music analytics, with important implications for artists, record labels, streaming platforms, and music recommendation systems.

Recent advances in machine learning have substantially improved the ability to predict song popularity using measurable audio characteristics. Spotify audio features, including danceability, energy, loudness, acousticalness, instrumentality, and speechiness, have become widely adopted for modelling commercial success because they provide standardized representations of musical content. Previous studies consistently demonstrate that these features contain meaningful predictive information, with ensemble learning methods, particularly Random Forest, frequently outperforming conventional linear models in hit-song prediction tasks [2], [3], [4]. At the same time, the increasing adoption of Explainable Artificial Intelligence (XAI) has shifted attention beyond predictive performance toward

*Corresponding author: Christopher Andreas (christopher.andreas@ciputra.ac.id)

DOI: <https://doi.org/10.47738/jads.v7i4.1506>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

understanding how machine learning models derive their predictions. Among existing XAI techniques, SHAP has emerged as one of the most widely adopted approaches for quantifying feature contributions and interpreting complex predictive models [5], [6].

Although previous studies demonstrate that hit-song prediction is feasible, they generally formulate the problem as a static classification task by training a single predictive model on data collected across multiple years. Such an approach implicitly assumes that the relationships between audio characteristics and commercial success remain stable over time. However, the recorded-music industry has undergone profound technological and cultural transformations during the past six decades, including the transition from the Analog era to the CD/Digital era and subsequently to the Streaming era [7]. These transitions represent more than changes in distribution technology. They coincide with major shifts in music-production practices, consumer listening behaviour, market competition, recommendation mechanisms, and the economic structure of the recording industry. The transition from physical media to digital distribution substantially reduced production and distribution barriers, while streaming platforms increasingly relied on algorithmic recommendation systems and data-driven content discovery, fundamentally altering how audiences' access and consume music. At the same time, evolving production practices, including the widespread adoption of digital audio workstations, advances in mastering techniques, and the so-called "loudness war", have progressively changed the acoustic characteristics of commercially successful recordings [8], [9], [10]. Consequently, the audio attributes associated with commercial success may reasonably be expected to evolve alongside these technological, industrial, and cultural transformations rather than remain temporally invariant.

Furthermore, commercial success itself is a dynamic concept. Throughout the study period, the methodology used to compile the Billboard Hot 100 evolved from being dominated by physical sales and radio airplay to incorporating digital downloads and, more recently, streaming activity. Although the Billboard Hot 100 remains one of the most widely adopted indicators of commercial success in music analytics, these methodological changes imply that the operational definition of a "hit song" has also evolved over time. Therefore, temporal comparisons should be interpreted as reflecting changes in songs that achieved commercial success under the prevailing market conditions of each historical period rather than assuming a perfectly invariant definition across six decades [11]. Consequently, studies that focus primarily on predictive accuracy provide only limited understanding of how the importance of musical attributes changes across different historical periods.

To address this limitation, this study adopts a temporal perspective on hit-song prediction by investigating the evolution of commercially successful songs across three music-industry eras spanning 1960 to 2019. The three-era segmentation (Analog, CD/Digital, and Streaming) is adopted because each period reflects a distinct technological and industrial environment characterized by different production workflows, distribution channels, consumption behaviour, and market dynamics. Rather than representing arbitrary chronological intervals, these eras capture major transitions in the recorded-music ecosystem that are expected to influence the relationship between intrinsic audio characteristics and commercial success. Rather than constructing a single global prediction model, separate machine learning models are developed for each era to evaluate whether predictive performance and the relative importance of Spotify audio features remain consistent over time. Logistic Regression, Random Forest, and K-Nearest Neighbours (KNN) are employed as representative linear, ensemble, and instance-based learning algorithms, while SHAP is used to quantify and compare feature importance across historical periods [12], [13]. Because the objective of this study is to quantify the predictive contribution of intrinsic audio characteristics, external determinants of commercial success, such as artist popularity, marketing expenditure, playlist placement, release timing, record-label support, and audience engagement, are intentionally excluded from the modelling framework. Accordingly, the resulting models should be interpreted as explaining the contribution of audio features to hit-song prediction rather than providing a complete explanation of all factors influencing commercial success. In addition, multiple robustness analyses, including repeated random seeds, bootstrap confidence intervals for cross-era SHAP differences, permutation importance, alternative era definitions, and data de-duplication, are conducted to assess the stability of the observed explanation patterns.

Unlike previous studies that primarily emphasize prediction accuracy, this study investigates how the determinants of commercial success evolve over time through the integration of predictive modelling and Explainable Artificial Intelligence. Specifically, the study makes four main contributions. First, it reframes hit-song prediction as a temporal analytical problem by examining the evolution of commercially successful songs across distinct music-industry eras rather than assuming a static relationship between audio characteristics and chart performance. Second, it introduces a temporally stratified analytical framework that enables systematic comparison of predictive performance and feature importance across historical periods. Third, it integrates SHAP-based Explainable Artificial Intelligence with machine learning to characterize the temporal evolution of Spotify audio feature importance, providing interpretable evidence

of how the audio signatures of successful songs have changed over six decades. Finally, it evaluates both the robustness and statistical reliability of the observed temporal patterns through complementary validation procedures, including repeated random seeds, bootstrap confidence intervals for cross-era differences in SHAP importance, permutation importance, alternative era definitions, and data de-duplication, thereby strengthening confidence that the reported temporal evolution reflects consistent predictive patterns rather than model-specific variability.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on hit-song prediction, machine learning, and Explainable Artificial Intelligence. Section 3 presents the dataset and research methodology, model development, temporal explainability, and robustness assessment. Section 4 reports the experimental results, explainability analyses, and discusses their implications and limitations. Section 5 concludes the paper and outlines directions for future research.

2. Literature Review

2.1. Hit Song Prediction: Current Progress and Research Gap

Predicting commercially successful songs has become an active research topic in music analytics owing to the availability of large-scale streaming datasets and standardized audio descriptors. Previous studies consistently demonstrate that Spotify audio features provide informative representations for modelling commercial success and that machine learning algorithms can effectively distinguish hit songs from non-hit songs [2], [3], [4]. Collectively, these findings establish a strong foundation for data-driven hit-song prediction.

Early work by Herremans et al. [3] demonstrated the feasibility of predicting dance-chart hits using machine learning models trained on audio features, showing that a relatively small subset of musical attributes contains substantial predictive information. Interiano et al. [2] substantially extended this research by analyzing more than 500,000 songs released between 1985 and 2015. Their study showed that Random Forest achieved strong predictive performance while revealing long-term trends in musical characteristics. More recently, Saragih [4] evaluated several machine learning algorithms using Spotify audio features from Indonesian streaming data and reported that Random Forest consistently outperformed competing methods, while louder and less instrumental songs were generally associated with higher popularity. Table 1 summarizes the methodological characteristics of representative studies on hit-song prediction and highlights the position of the present research within the existing literature.

Despite these encouraging results, previous studies also acknowledge that commercial success is influenced by multiple external factors beyond intrinsic audio characteristics, including artist popularity, marketing activities, playlist exposure, release timing, record-label support, and audience engagement. Consequently, machine learning models based solely on audio features should be interpreted as capturing only the predictive contribution of musical attributes rather than the complete set of determinants underlying commercial success. As shown in table 1, previous studies consistently demonstrate the effectiveness of machine learning for predicting commercially successful songs.

Table 1. Comparison of Previous Studies on Hit Song Prediction and the Research Gap Addressed in This Study

Study	Dataset	Machine Learning	Explainable AI	Temporal Perspective	Robustness Analysis	Main Limitation
Herremans et al. [3]	Dance chart songs	Classification models	No	No	No	Focused primarily on prediction performance.
Interiano et al. [2]	>500,000 songs (1985–2015)	Random Forest	No	Long-term trend analysis	No	Did not compare feature importance across distinct musical eras.
Saragih [4]	Indonesian Spotify dataset	Random Forest and other ML models	No	No	No	Limited to a contemporary streaming dataset.
This study	Billboard-labelled Spotify songs (1960–2019)	Logistic Regression, Random Forest, KNN	SHAP	Three music-industry eras	Multiple robustness analyses	—

Nevertheless, four important limitations remain. First, hit-song prediction has largely been formulated as a static classification problem, assuming that the relationships between audio characteristics and commercial success remain unchanged over time. Second, although temporal trends in musical characteristics have been reported, the evolution of

feature importance across clearly defined music-industry eras has received little systematic investigation [2]. Third, existing studies primarily emphasize predictive performance, with limited attention to the robustness of explanation-based findings. Finally, relatively little attention has been devoted to assessing whether the observed temporal differences in feature importance are statistically robust across independently trained models. The present study addresses these limitations by integrating temporally stratified machine learning models with SHAP-based Explainable Artificial Intelligence to investigate the temporal evolution of hit-song predictors and evaluate the stability of the resulting explanations.

2.2. Machine Learning Paradigms for Hit Song Prediction

Machine learning has become a standard approach for modelling commercial music success because it enables complex relationships between audio characteristics and popularity to be learned directly from data. For structured tabular datasets such as Spotify audio features, classical machine learning algorithms remain widely adopted owing to their balance between predictive performance, computational efficiency, and interpretability [2], [3], [4], [14]. Commercial success, however, is influenced by both intrinsic musical characteristics and external market factors, including artist popularity, marketing activities, playlist exposure, release timing, record-label support, and audience engagement [15]. Many of these variables are unavailable or inconsistently measured over long historical periods, making them difficult to incorporate into longitudinal analyses spanning multiple decades. Consequently, the present study intentionally focuses on intrinsic audio characteristics represented by Spotify audio features in order to isolate their predictive contribution while acknowledging that the resulting models do not aim to explain the full complexity of commercial success.

Rather than identifying the state-of-the-art predictive algorithm, this study compares three representative learning paradigms that capture different assumptions regarding the relationship between predictor variables and commercial success. Logistic Regression serves as a linear probabilistic baseline, Random Forest models nonlinear interactions through ensemble learning, and K-Nearest Neighbours (KNN) performs classification based on local similarity in the feature space [14], [16], [17]. These complementary algorithms were selected because they provide a balance between predictive performance, computational efficiency, and interpretability, thereby facilitating consistent comparisons of temporal patterns across music-industry eras rather than maximizing predictive accuracy through increasingly complex model architectures.

Random Forest has consistently demonstrated strong predictive performance in previous music analytics studies [2], [3], [4]. More importantly, its tree-based structure allows efficient interpretation through TreeSHAP, providing exact feature-attribution estimates for analyzing temporal changes in feature importance [6], [18]. Consequently, the selected models are intended not only to compare predictive performance but also to establish a complementary analytical framework for the explainability analysis presented in the following section. Accordingly, the emphasis of this study is on understanding how the predictive contribution of intrinsic audio characteristics evolves over time, rather than on constructing a comprehensive model of all determinants of commercial success.

2.3. Explainable Artificial Intelligence Using SHAP

Machine learning models often achieve high predictive performance at the expense of interpretability, particularly when nonlinear models are employed. Consequently, Explainable Artificial Intelligence (XAI) has become an important component of predictive analytics by enabling researchers to understand how input variables contribute to model predictions [5], [6]. Such interpretability is particularly valuable when the objective extends beyond prediction to understanding the mechanisms underlying observed patterns.

Among existing XAI methods, SHAP (SHapley Additive exPlanations) has become one of the most widely adopted approaches because it provides a theoretically grounded and consistent measure of feature contribution based on cooperative game theory. For tree-based models, TreeSHAP enables efficient and exact computation of Shapley values, making it well suited for interpreting Random Forest models [19].

In contrast to previous hit-song prediction studies that primarily emphasize predictive performance, this study employs SHAP to compare feature importance across independently trained models representing different music-industry eras. This approach enables the investigation of whether the audio characteristics associated with commercial success remain stable or evolve over time. Beyond descriptive comparison, the temporal evolution of feature importance is further evaluated through bootstrap confidence intervals for cross-era differences in mean absolute SHAP values, enabling

statistical assessment of whether the observed changes exceed sampling variability. To improve the reliability of these findings, the SHAP analysis is complemented by repeated random seeds, permutation importance, alternative era definitions, and data de-duplication. Together, these analyses provide a rigorous basis for interpreting temporal changes in hit-song predictors.

3. Methodology

3.1. Research Design and Dataset

This study adopts a comparative machine learning framework to investigate how the audio characteristics associated with commercially successful songs evolve over time. Rather than developing a single predictive model for the entire dataset, separate models are constructed for different music-industry eras to examine whether the determinants of chart success remain stable under changing technological and cultural contexts. The overall research workflow comprises four sequential stages: data preparation, model development, explainability analysis, and robustness assessment.

The analysis is based on the Spotify Hit Predictor Dataset (1960–2019), which contains 41,106 Billboard-labelled tracks described by Spotify audio features [20]. Each track is labelled as either a *hit*, indicating that it reached the Billboard Hot 100, or a *flop* otherwise. The non-hit tracks were constructed as class-balanced control observations within each decade file of the original dataset, resulting in an equal number of hit and flop songs for every decade. Billboard chart appearance was adopted as an operational definition of commercial success because it provides a consistent and widely used benchmark in music analytics. Nevertheless, Billboard chart methodology has evolved over time through changes in the relative contribution of physical sales, radio airplay, digital downloads, and streaming activity. Consequently, the findings should be interpreted as reflecting the evolution of audio characteristics associated with Billboard-defined commercial success rather than an absolute measure of musical popularity across historical periods. Following data quality screening, three observations containing invalid tempo or time-signature values were removed, resulting in a final dataset of 41,103 tracks.

Fifteen numerical Spotify audio features were retained as predictor variables: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentality, liveness, valence, tempo, duration, time signature, chorus-hit position, and number of sections. These variables capture complementary aspects of musical content, including rhythm, harmony, dynamics, timbre, and song structure. Identifier-related attributes, including track name, artist name, and record identifiers, were excluded because they do not represent intrinsic audio characteristics and therefore provide limited value for developing generalizable prediction models. The present study intentionally focuses on intrinsic audio characteristics. External determinants of commercial success, such as artist popularity, marketing activities, playlist placement, release timing, record-label support, social-media exposure, and audience engagement, were not incorporated because comparable historical measurements spanning six decades are unavailable. Accordingly, the developed models quantify the predictive contribution of audio characteristics rather than the complete set of factors influencing commercial success.

To investigate temporal changes in the determinants of commercial success, the dataset was partitioned into three music-industry eras representing major transitions in music production, distribution, and consumption: Analog (1960–1989; $n = 23,315$), CD/Digital (1990–2009; $n = 11,391$), and Streaming (2010–2019; $n = 6,397$). This temporal stratification was motivated not only by technological transitions in music distribution but also by substantial changes in production practices, market structure, and music-consumption behaviour across the recorded-music industry. Grouping the data into these three eras provides a balance between historical interpretability and sufficient sample size for reliable model estimation while enabling meaningful comparison across distinct industry environments. In addition to the era-level analysis, decade-specific models were also developed to provide a finer-grained assessment of long-term temporal trends.

Prior to model development, the dataset underwent exploratory analysis. Descriptive statistics, class-wise feature distributions, and correlation analysis were used to characterize the data and identify potential relationships among predictor variables. Continuous features were subsequently standardized where required by the learning algorithm to ensure comparable feature scales. All preprocessing, model training, and subsequent analyses were conducted independently within each era, ensuring that predictive performance and feature-importance estimates reflected era-specific characteristics rather than patterns learned from pooled observations across different historical periods. This analytical design enabled fair temporal comparisons while avoiding information leakage across different historical periods.

3.2. Model Development

To evaluate the temporal predictability of commercially successful songs, three supervised machine learning algorithms representing complementary modelling paradigms were considered: Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbours (KNN). Logistic Regression was selected as a linear baseline model because of its simplicity and interpretability, Random Forest was included as a nonlinear ensemble learning method capable of modelling complex feature interactions, and KNN was employed as a representative instance-based classifier. Together, these models provide a balanced comparison across linear, tree-based, and distance-based learning approaches that are widely adopted in predictive analytics [21].

Model development was performed independently for each music-industry era to ensure that predictive relationships reflected the characteristics of the corresponding historical period. For algorithms sensitive to feature magnitude, namely Logistic Regression and KNN, continuous variables were standardized using z-score normalization based exclusively on the training data to prevent information leakage [22]. Random Forest was trained using the original feature values because tree-based methods are invariant to monotonic feature scaling.

Rather than adopting manually selected model settings, hyperparameters were optimized using exhaustive grid-search combined with three-fold stratified cross-validation, with the mean ROC–AUC used as the optimization criterion. The search space included the regularization strength and penalty type for Logistic Regression, the number of trees and maximum tree depth for Random Forest, and the number of neighbours together with the weighting strategy for KNN. The complete hyperparameter search space is summarized in table 2. Hyperparameter optimization was performed independently for each music-industry era while maintaining an identical search strategy across all experiments, thereby ensuring a fair comparison of predictive performance under consistent model-selection procedures [23].

Table 2. Hyperparameter Search Space

Model	Hyperparameters
Logistic Regression	$C = \{0.01, 0.1, 1, 10, 100\}$; $\text{penalty} = \{L1, L2\}$
Random Forest	$n_estimators = \{150, 300\}$; $\text{max_depth} = \{12, 20\}$
KNN	$n_neighbors = \{15, 45, 75\}$; $\text{weights} = \{\text{uniform}, \text{distance}\}$

All experiments were implemented in Python 3.11 using the Scikit-learn library on an Apple M3 workstation with 16 GB unified memory. Logistic Regression employed the liblinear solver with a maximum of 4,000 iterations, whereas Random Forest and K-Nearest Neighbours used the standard Scikit-learn implementations. Fixed random seeds were applied throughout model fitting and cross-validation to ensure reproducible results, and parallel computation was enabled where applicable using Joblib.

The predictive performance of the optimized models was subsequently evaluated using Accuracy, F1-score, and ROC–AUC. Accuracy provides an overall measure of classification performance, F1-score balances precision and recall under class prediction, whereas ROC–AUC evaluates the ability of the classifier to discriminate between hit and flop songs across all decision thresholds. Because the dataset is class-balanced by construction, these metrics collectively provide a comprehensive assessment of classification performance while facilitating direct comparison across different music-industry eras.

3.3. Temporal Explainability and Robustness Assessment

To characterize how the determinants of commercial success evolve across different historical periods, this study combines model explainability with a comprehensive robustness assessment to enable reliable comparison of feature importance between independently trained models. Rather than focusing solely on predictive performance, the proposed framework examines whether the contribution of Spotify audio features remains stable or changes across music-industry eras, thereby providing insight into the long-term evolution of hit-song characteristics.

To interpret model behaviour, SHapley Additive exPlanations (SHAP) were employed to quantify the contribution of individual audio features to model predictions [5], [6], [18]. SHAP values were computed using TreeExplainer, which provides exact and computationally efficient Shapley values for tree-based models. Global feature importance was quantified using the mean absolute SHAP value for each predictor, representing the average magnitude of its contribution irrespective of direction. For feature j , global SHAP importance is calculated as shown in Equation 1.

$$I_j = \frac{1}{N} \sum_{i=1}^N |\phi_{ij}| \quad (1)$$

I_j denotes the global importance of feature j , N is the number of observations evaluated within an era, and ϕ_{ij} represents the absolute SHAP value of feature j for observation i . This measure enables consistent comparison of feature importance across independently trained era-specific Random Forest models. To complement the global feature-importance analysis, the direction of association between each predictor and hit-song classification was examined using standardised Logistic Regression coefficients, where positive coefficients indicate an increased likelihood of the hit class and negative coefficients indicate the opposite. To quantify temporal changes in feature importance, the difference in mean absolute SHAP importance between two music-industry eras (a) and (b) is calculated as shown in Equation 2.

$$\Delta SHAP_j^{(a,b)} = I_j^{(a)} - I_j^{(b)} \quad (2)$$

$I_j^{(a)}$ and $I_j^{(b)}$ denote the mean absolute SHAP importance of feature j in eras (a) and (b), respectively. A positive $\Delta SHAP_j^{(a,b)}$ indicates that feature j has greater predictive contribution in era (a), whereas a negative value indicates greater importance in era (b). This measure was used to identify whether the relative contribution of individual audio features increased or decreased across successive music-industry eras.

To evaluate whether the observed temporal changes in feature importance represent meaningful differences rather than sampling variation, bootstrap resampling with 1,000 stratified replicates per era was additionally performed. For each bootstrap replicate, the corresponding difference in mean absolute SHAP importance was recalculated to obtain an empirical distribution of $\Delta SHAP$. The 95% confidence interval was estimated using the percentile bootstrap method, as shown in Equation 3.

$$CI_{95\%} = [Q_{0,025}(\Delta SHAP_j), Q_{0,975}(\Delta SHAP_j)] \quad (3)$$

$Q_{0,025}$ and $Q_{0,975}$ denote the 2.5th and 97.5th percentiles of the bootstrap distribution. Temporal differences were considered statistically significant when the corresponding 95% confidence interval excluded zero.

The stability of the observed temporal patterns was evaluated from multiple complementary perspectives. First, model stability was assessed by repeating the SHAP analysis across five independent random seeds, with 95% bootstrap confidence intervals estimated for all feature-importance scores. Second, predictive superiority was examined using paired bootstrap resampling (2,000 iterations) to determine whether differences in ROC-AUC between Random Forest and the competing classifiers were statistically significant. Third, methodological robustness was evaluated through a series of sensitivity analyses, including alternative era definitions, removal of duplicated tracks appearing across decade files, and comparison between SHAP-based feature importance and model-agnostic permutation importance.

Collectively, these procedures establish a robust analytical framework for investigating the temporal evolution of hit-song predictors. By integrating explainability with multiple robustness assessments, formal statistical comparison of cross-era SHAP importance, and sensitivity analyses, the proposed methodology reduces the likelihood that the observed changes in feature importance arise from model instability, sampling variation, or arbitrary data partitioning, thereby providing more reliable evidence of long-term changes in the audio characteristics associated with commercially successful songs. The overall analytical framework is illustrated in [figure 1](#), while the corresponding procedural steps are provided in [algorithm 1](#).

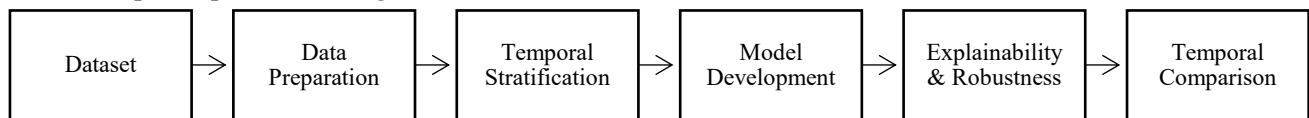


Figure 1. Overall analytical framework for temporal explainable machine learning of hit-song predictors.

Algorithm 1. Temporal Explainable Machine Learning and Robustness Assessment Framework

Input: Dataset D , era definitions E , models $\{LR, RF, KNN\}$

- 1: Load and clean D ; retain 15 Spotify audio features
 - 2: Perform exploratory data analysis
 - 3: Partition D into music-industry eras E
 - 4: for each era $e \in E$ do
 - 5: Split data; standardize features for LR and KNN using training-set parameters only
 - 6: Optimize LR, RF, and KNN using grid search with 3-fold stratified cross-validation
 - 7: Train optimized models and evaluate Accuracy, F1-score, and ROC-AUC
 - 8: Compute RF SHAP values and standardized LR coefficients
 - 9: end for
-

- 10: Compare RF feature importance across eras using stratified bootstrap Δ SHAP analysis ($B = 1,000$)
11: Assess robustness using multiple seeds, paired bootstrap ($B = 2,000$), and sensitivity analyses
12: Report temporal changes in predictive performance and feature importance
Output: Era-specific model performance, feature importance, temporal patterns, and robustness results

4. Results and Discussion

4.1. Exploratory Data Analysis

Prior to predictive modelling, exploratory data analysis (EDA) was conducted to examine the overall characteristics of the dataset and to identify preliminary differences between hit and flop songs. Table 3 summarizes the descriptive statistics of the fifteen Spotify audio features across all 41,103 tracks, while figure 2 illustrates the distributions of the eight most informative features for the two classes.

Table 3. Descriptive statistics for the fifteen audio features (all 41,103 tracks).

Feature	Mean	SD	Min	Median	Max
danceability	0.54	0.178	0.058	0.552	0.988
energy	0.58	0.253	0	0.601	1
key	5.213	3.535	0	5	11
loudness	-10.221	5.311	-49.253	-9.257	3.744
mode	0.693	0.461	0	1	1
speechiness	0.073	0.086	0.022	0.043	0.96
acousticness	0.364	0.339	0	0.258	0.996
instrumentalness	0.154	0.304	0	0	1
liveness	0.202	0.173	0.013	0.132	0.999
valence	0.542	0.267	0	0.559	0.996
tempo	119.343	29.093	31.988	117.567	241.423
duration_ms	234880.263	118971.258	15168	217907	4170227
time_signature	3.894	0.422	1	4	5
chorus_hit	40.104	18.999	0	35.85	433.182
sections	10.476	4.872	0	10	169

Several audio characteristics exhibit clear separation between commercially successful and unsuccessful songs. Compared with flops, hit songs generally have higher loudness, lower instrumentalness, lower acousticness, and higher valence. These differences are reflected in the class distributions shown in figure 2 and suggest that commercially successful songs are generally associated with higher loudness, lower instrumentalness, lower acousticness, and higher valence within the present dataset. In contrast, features such as key, mode, and time signature display substantial overlap between the two classes, indicating relatively limited discriminative power when considered individually.

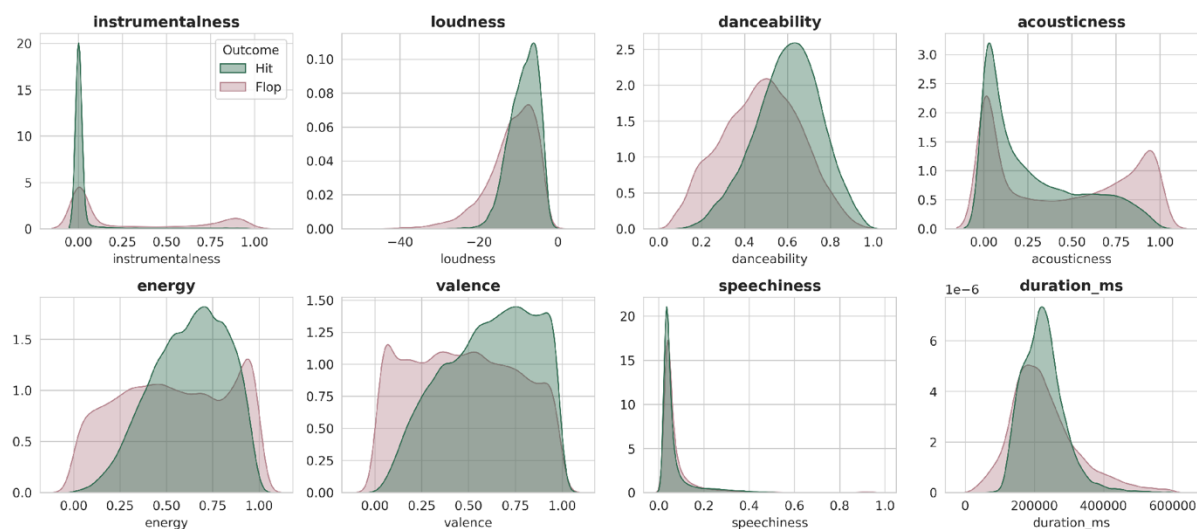


Figure 2. Distributions of eight key features for hits versus flops.

However, these observations should be interpreted as descriptive associations rather than evidence that individual audio characteristics independently determine commercial success. The distributions presented in [figure 2](#) reflect only intrinsic musical attributes captured by Spotify audio features, whereas real-world commercial performance is also influenced by numerous external factors, including artist popularity, marketing activities, playlist placement, release timing, audience engagement, record-label support, and broader music-industry dynamics, none of which are represented in the present dataset. Consequently, the subsequent analyses evaluate the predictive contribution of audio characteristics rather than the complete set of determinants underlying hit-song success.

Overall, the exploratory analysis indicates that the selected predictors provide sufficient variability and class discrimination to support subsequent machine learning modelling. At the same time, the descriptive statistics suggest that the relationship between audio characteristics and commercial success is multidimensional, motivating the use of machine learning models capable of capturing both linear and nonlinear interactions among features. The following section therefore evaluates whether these multivariate relationships can accurately distinguish hit songs from flops across different music-industry eras.

4.2. Predictive Performance

Hyperparameter optimization using grid-search cross-validation produced consistent model configurations across the three music-industry eras. Random Forest repeatedly selected 300 trees with a maximum depth of 20, whereas K-Nearest Neighbours consistently favoured distance-based weighting with relatively large neighbourhood sizes (45–75 neighbours). Logistic Regression required stronger regularization in the Analog era than in the subsequent eras. The selected hyperparameters are summarized in [table 4](#).

Table 4. Hyperparameters selected by grid search (3-fold cross-validation) for each model and era.

Era	Classifier	Best hyperparameters	ROC-AUC
Analog (1960–89)	Logistic Regression	C = 1, L1 penalty	0.786
	Random Forest	300 trees, max_depth = 20	0.848
	KNN	k = 75, distance-weighted	0.810
CD/Digital (1990–2009)	Logistic Regression	C = 100, L2 penalty	0.874
	Random Forest	300 trees, max_depth = 20	0.917
	KNN	k = 45, distance-weighted	0.889
Streaming (2010s)	Logistic Regression	C = 100, L2 penalty	0.864
	Random Forest	300 trees, max_depth = 20	0.912
	KNN	k = 75, distance-weighted	0.863

The predictive performance of the optimized models is presented in [table 5](#), while [figure 3](#) illustrates their comparative performance across eras. Among the evaluated algorithms, Random Forest consistently achieved the highest predictive performance, outperforming both Logistic Regression and K-Nearest Neighbours for Accuracy, F1-score, and ROC–AUC in every era. The highest ROC–AUC values reached 0.848 for the Analog era, 0.917 for the CD/Digital era, and 0.912 for the Streaming era, demonstrating that nonlinear ensemble learning is more effective at capturing the complex relationships between Spotify audio features and commercial success than either linear or distance-based approaches.

Table 5. Cross-validated performance of the tuned classifiers by era.

Era	Classifier	Accuracy	F1	ROC-AUC
Analog (1960–89)	Logistic Regression	0.713	0.734	0.786
	Random Forest	0.769	0.782	0.848
	KNN	0.723	0.764	0.810
CD/Digital (1990–2009)	Logistic Regression	0.800	0.810	0.874
	Random Forest	0.846	0.851	0.917
	KNN	0.797	0.823	0.889
Streaming (2010s)	Logistic Regression	0.804	0.819	0.864
	Random Forest	0.843	0.850	0.912
	KNN	0.769	0.807	0.863

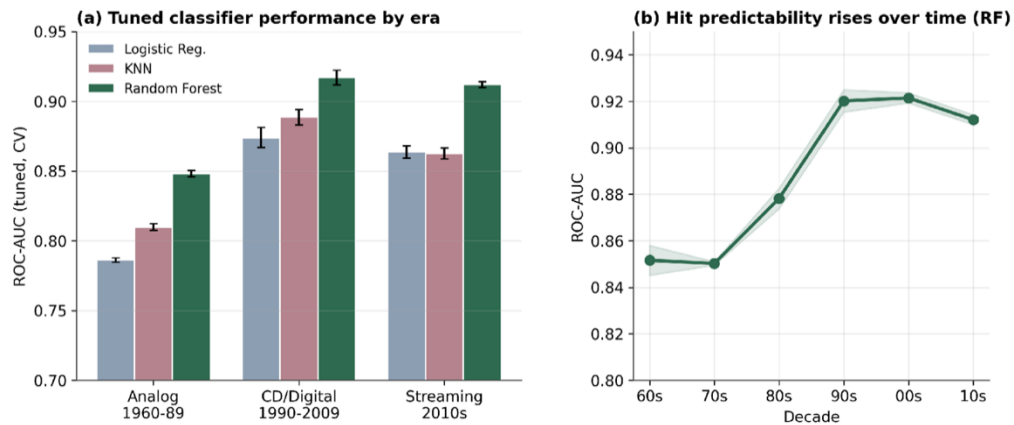


Figure 3. Tuned classifier performance by era (left) and the rise in hit predictability across decades for the Random Forest (right).

Beyond the comparison between classifiers, an important temporal pattern emerges. Predictive performance improves from the Analog era to the CD/Digital era and remains consistently high during the Streaming era. At the decade level, Random Forest exhibits a gradual increase in ROC–AUC from approximately 0.85 in the 1960s and 1970s to around 0.92 during the 2000s. This finding indicates that Spotify audio characteristics become increasingly informative for discriminating between hit and flop songs across successive historical periods within the present dataset. However, this improvement should not be interpreted as evidence that commercial success itself has become intrinsically more predictable. Several factors may contribute to the observed increase in predictive performance, including increasing standardization of music production, advances in recording and mastering practices, changes in Billboard chart compilation methodology, shifts in listener behaviour and platform-mediated music discovery, or reduced diversity among commercially successful songs. Consequently, the observed performance trend is more appropriately interpreted as reflecting changes in the predictive relationship between Spotify audio features and Billboard chart outcomes rather than a single underlying causal mechanism.

These findings are broadly consistent with previous studies reporting the strong predictive capability of Random Forest for modelling music popularity [2], [3], [4]. However, whereas earlier work primarily emphasized prediction accuracy, the present results establish the foundation for the subsequent explainability analysis by demonstrating that era-specific models provide reliable predictive performance across all historical periods. The following section therefore investigates whether the audio characteristics driving these predictions also evolve over time.

4.3. Temporal Evolution of Feature Importance

Having established that the Random Forest models provide reliable predictive performance across all music-industry eras, the next analysis examines whether the audio characteristics underlying these predictions remain stable over time. Table 6 presents the mean absolute SHAP values for each feature across the three eras, while figure 4 and figure 5 illustrate the corresponding importance rankings and decade-level trajectories.

Table 6. Mean absolute SHAP value (and importance rank) per feature in each era, ordered by Analog-era importance.

Feature	Analog (1960–89)	CD/Digital (1990–2009)	Streaming (2010s)
instrumentalness	0.095 (#1)	0.113 (#1)	0.169 (#1)
acousticness	0.056 (#2)	0.065 (#3)	0.047 (#4)
speechiness	0.051 (#3)	0.025 (#7)	0.009 (#8)
danceability	0.047 (#4)	0.090 (#2)	0.059 (#3)
loudness	0.030 (#5)	0.038 (#5)	0.063 (#2)
energy	0.029 (#6)	0.027 (#6)	0.042 (#5)
valence	0.025 (#7)	0.015 (#8)	0.020 (#7)
duration_ms	0.019 (#8)	0.065 (#4)	0.031 (#6)
mode	0.018 (#9)	0.004 (#12)	0.001 (#15)
sections	0.012 (#10)	0.013 (#9)	0.007 (#9)
tempo	0.011 (#11)	0.008 (#11)	0.006 (#10)
liveness	0.009 (#12)	0.008 (#10)	0.006 (#11)

time_signature	0.004 (#13)	0.002 (#15)	0.004 (#13)
chorus_hit	0.003 (#14)	0.004 (#13)	0.004 (#12)
key	0.003 (#15)	0.003 (#14)	0.003 (#14)

A clear pattern of temporal evolution is observed. Instrumentalness remains the most influential predictor throughout the entire study period, with its mean absolute SHAP value increasing from 0.095 in the Analog era to 0.169 in the Streaming era. This steady increase indicates that the distinction between vocal and instrumental music has become progressively more important in differentiating commercially successful songs from flops. In contrast, loudness exhibits the largest upward shift in relative importance, rising from the fifth most influential feature in the Analog era to the second position during the Streaming era. The growing contribution of loudness suggests that production characteristics have become increasingly associated with chart success. These observations are based on global SHAP importance estimates and are subsequently evaluated using bootstrap confidence intervals to determine whether the observed cross-era differences exceed sampling variability.

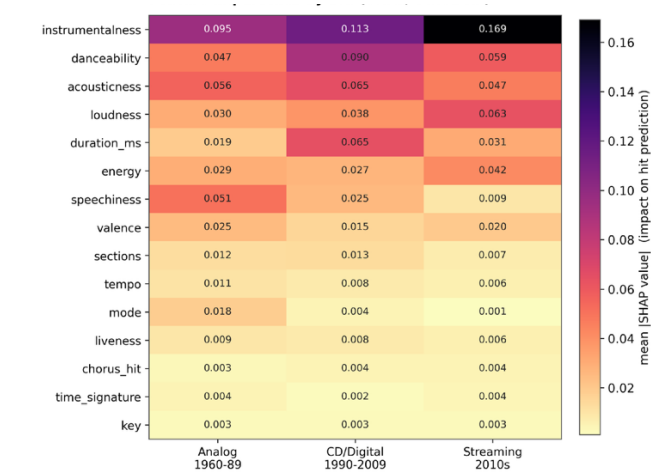


Figure 4. Feature importance (mean |SHAP|) by era for the tuned Random Forest.

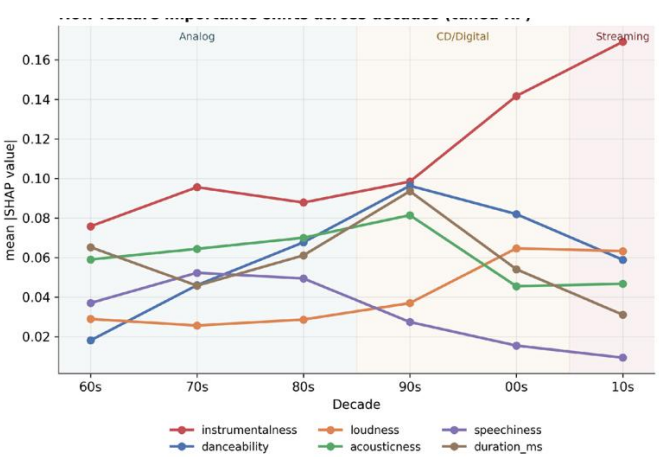


Figure 5. Feature-importance trajectories (mean |SHAP|) across the six decades.

Not all features display increasing importance. Speechiness shows a substantial decline across eras, dropping from the third most important predictor in the Analog era to eighth in the Streaming era. Acousticness also becomes less influential overall, although the subsequent bootstrap analysis indicates that most of this reduction occurs during the transition from the CD/Digital era to the Streaming era. By contrast, danceability reaches its highest importance during the CD/Digital era before declining slightly in the Streaming era, suggesting that its contribution varies with changes in musical styles rather than following a monotonic trend. Most of the remaining variables, including key, mode, tempo, time signature, and chorus-hit location, consistently exhibit relatively low SHAP values, indicating limited influence on model predictions.

The decade-level analysis shown in figure 4 further demonstrates that these changes occur gradually rather than abruptly. Instead of isolated fluctuations between adjacent decades, the principal predictors follow smooth trajectories over the six decades considered. The increasing importance of instrumentalness and loudness, together with the declining influence of speechiness and acousticness, therefore appears to reflect long-term changes in the audio characteristics associated with commercially successful songs rather than short-term variation within individual decades. These findings indicate that the determinants of hit-song success are not temporally stationary. While the predictive capability of Spotify audio features remains consistently high, the relative contribution of individual features evolves across historical periods. The following section examines the direction of these effects and discusses their implications for understanding long-term changes in music production and listener preferences. To determine whether these observed temporal differences represent genuine changes rather than sampling variability, pairwise comparisons of mean absolute SHAP values were conducted using the full-pipeline bootstrap procedure described in Section 3.3. Table 7 summarizes the estimated cross-era differences (Δ SHAP) and the corresponding Benjamini–Hochberg adjusted p-values.

Table 7. Statistical significance of pairwise changes in SHAP feature importance across music-industry eras based on full-pipeline bootstrap (B = 1,000). Positive values indicate increased importance in the later era.

Feature	Analog → CD/Digital ΔSHAP	p-value	CD/Digital → Streaming ΔSHAP	p-value	Analog → Streaming ΔSHAP	p-value
Instrumentalness	+0.0181	0.005*	+0.0557	0.005*	+0.0738	0.006*
Acousticness	+0.0096	0.226	−0.0186	0.009*	−0.0090	0.109
Speechiness	−0.0251	0.005*	−0.0161	0.005*	−0.0412	0.006*
Danceability	+0.0425	0.005*	−0.0310	0.005*	+0.0115	0.201
Loudness	+0.0081	0.060	+0.0250	0.005*	+0.0331	0.006*
Energy	−0.0025	0.965	+0.0149	0.005*	+0.0124	0.006*
Valence	−0.0094	0.005*	+0.0046	0.069	−0.0048	0.422
Duration_ms	+0.0453	0.005*	−0.0336	0.005*	+0.0117	0.043*
Mode	−0.0137	0.005*	−0.0034	0.040*	−0.0171	0.006*
Sections	+0.0010	0.927	−0.0062	0.030*	−0.0052	0.015*
Tempo	−0.0027	0.240	−0.0021	0.759	−0.0048	0.201
Liveness	−0.0011	0.927	−0.0023	0.486	−0.0034	0.369
Time_signature	−0.0026	0.217	+0.0019	0.169	−0.0006	0.791
Chorus_hit	+0.0012	0.965	+0.0002	0.238	+0.0014	0.204
Key	+0.0000	0.965	+0.0004	0.238	+0.0004	0.212

Table 7 confirms that the principal temporal patterns identified in table 6 are statistically supported. Instrumentalness increases significantly across all pairwise comparisons, whereas speechiness decreases significantly throughout the study period. Loudness becomes significantly more influential during the transition from the CD/Digital to the Streaming era, while danceability follows a non-monotonic trajectory, increasing significantly from the Analog to the CD/Digital era before declining significantly in the Streaming era, resulting in no significant net change between the first and last eras. Likewise, acousticness remains relatively stable during the first transition but declines significantly during the second transition, indicating that its reduced importance is primarily a recent phenomenon rather than a continuous long-term trend. Most of the remaining lower-ranked features show either small or statistically non-significant changes, suggesting that their contributions remain comparatively stable across historical periods.

While SHAP quantifies the magnitude of feature importance, it does not indicate whether higher feature values increase or decrease the likelihood of commercial success. To complement the SHAP analysis, figure 6 presents the standardized Logistic Regression coefficients estimated separately for each music-industry era, providing an interpretable summary of the direction of each feature's association with hit-song classification.

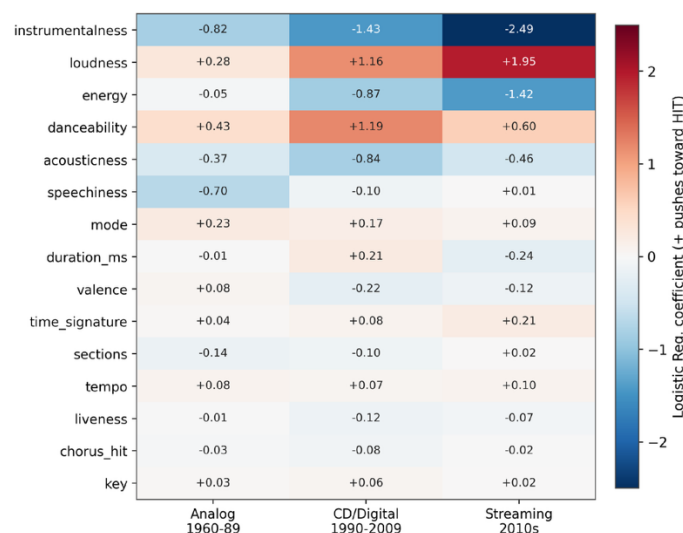


Figure 6. Direction of effect by era (standardized logistic-regression coefficients); positive values push toward the hit class.

The coefficient analysis reveals that the direction of the principal predictors remains broadly consistent across eras, although the strength of their effects changes substantially over time. Instrumentalness exhibits a consistently negative association with commercial success, with its standardized coefficient becoming increasingly negative from −0.82 in

the Analog era to -2.49 in the Streaming era. This trend indicates that songs containing fewer instrumental components and greater vocal presence have become progressively more likely to achieve chart success. Conversely, loudness displays an increasingly positive association with hit-song classification, with its coefficient rising from $+0.28$ to $+1.95$, suggesting that louder productions have become increasingly characteristic of commercially successful songs. Meanwhile, the influence of speechiness weakens across successive eras, consistent with its declining SHAP importance, whereas danceability exhibits its strongest positive association during the CD/Digital era before moderating in the Streaming era.

Taken together, the SHAP and coefficient analyses demonstrate that the determinants of commercial success have evolved not only in their relative importance but also in the strength of their associations with hit-song classification. Features related to vocal prominence and production intensity have become increasingly influential, whereas characteristics associated with acoustic and speech-oriented recordings contribute less to distinguishing successful songs in recent decades. These findings provide quantitative evidence that the audio signature of commercially successful music has undergone a gradual but systematic transformation over the past six decades.

4.4. Robustness Assessment and Sensitivity Analysis

To evaluate whether the observed temporal evolution of feature importance was robust to model variability, SHAP values were recomputed using five independent random seeds while maintaining the same Random Forest configuration as the primary analysis (figure 7).

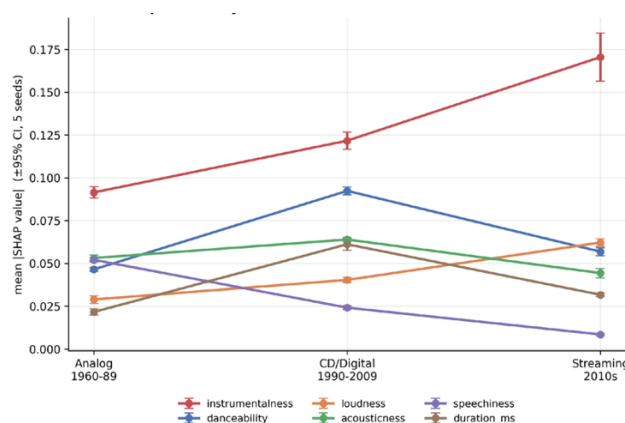


Figure 7. Feature importance by era with 95% confidence intervals over five random seeds.

Unlike the full-pipeline bootstrap analysis presented in Section 4.3, which evaluates statistical differences in feature importance across eras by resampling the underlying data, the present analysis keeps the dataset fixed and isolates variability arising from Random Forest initialization and model fitting. Figure 7 presents the corresponding mean feature-importance estimates together with their 95% bootstrap confidence intervals. Across all eras, the confidence intervals are narrow relative to the observed differences between eras, indicating that the temporal changes in feature importance are highly stable. In particular, the increasing importance of instrumentalness and loudness, together with the declining contribution of speechiness, remains consistent across all random seeds, suggesting that these trends are not attributable to stochastic variation during model training.

The superiority of Random Forest over the competing classifiers was further assessed using paired bootstrap resampling with 2,000 iterations on the pooled out-of-fold predictions. As summarized in table 8 and illustrated in figure 8 and figure 9, Random Forest significantly outperformed both Logistic Regression and K-Nearest Neighbours across all music-industry eras, with the 95% confidence intervals for the differences in ROC-AUC consistently excluding zero. These results provide statistical evidence that the observed performance advantage of Random Forest is unlikely to have arisen from sampling variability.

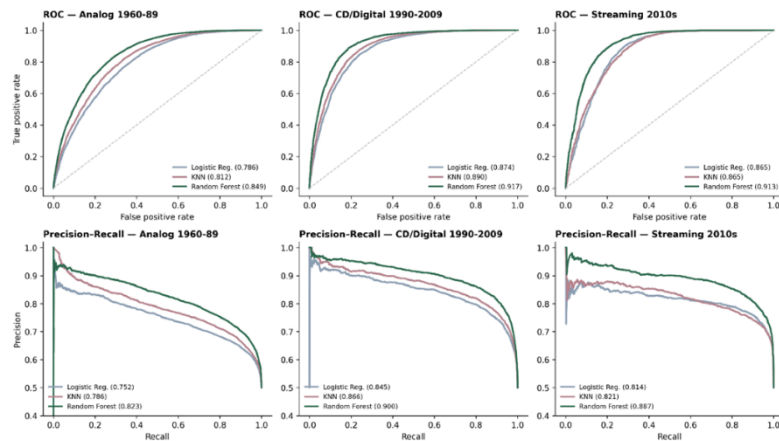


Figure 8. ROC and precision–recall curves by era from five-fold out-of-fold predictions.

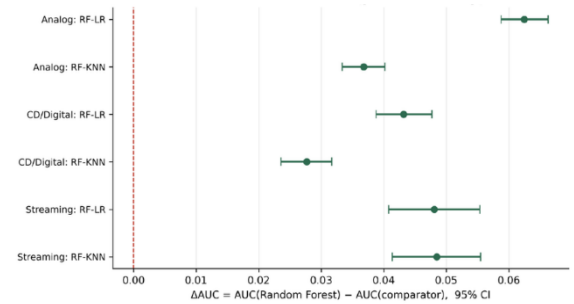


Figure 9. Random Forest versus the other classifiers: Δ AUC with 95% bootstrap confidence intervals (all above zero).

Table 8. Paired-bootstrap comparison of ROC-AUC (Random Forest minus each comparator) by era, 2,000 resamples.

Era	Comparison	Δ AUC	95% CI
Analog	RF – Logistic Reg.	+0.062	[+0.059, +0.066]
Analog	RF – KNN	+0.037	[+0.033, +0.040]
CD/Digital	RF – Logistic Reg.	+0.043	[+0.039, +0.048]
CD/Digital	RF – KNN	+0.028	[+0.024, +0.032]
Streaming	RF – Logistic Reg.	+0.048	[+0.041, +0.055]
Streaming	RF – KNN	+0.048	[+0.041, +0.056]

Finally, several sensitivity analyses were conducted to examine whether the principal findings depended on specific analytical choices (see [figure 10](#)). Repeating the analysis using alternative era definitions produced nearly identical temporal patterns of feature importance, while removing duplicated tracks appearing across decade files had negligible effects on both predictive performance and SHAP rankings. In addition, model-agnostic permutation importance exhibited strong agreement with the SHAP-based rankings (Spearman's $\rho = 0.94$ – 0.97), demonstrating that the identified temporal evolution is not specific to a single explanation technique. Collectively, these robustness analyses provide strong evidence that the reported changes in hit-song predictors represent stable characteristics of the data rather than artifacts of model specification, data partitioning, or feature-attribution methodology.

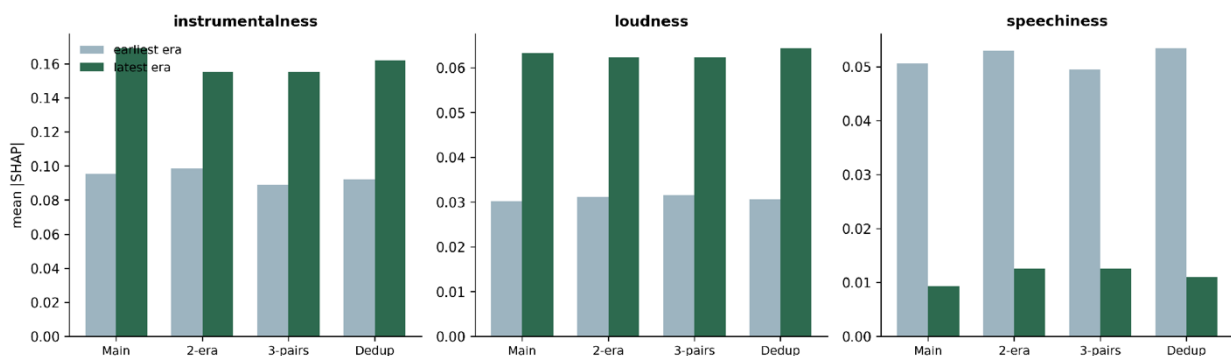


Figure 10. Robustness of the era shift: earliest versus latest era under the main, two-era, alternative three-era, and de-duplicated groupings.

4.5. Interpretation and Implications

The findings demonstrate that the audio characteristics associated with commercially successful songs are not temporally stationary but evolve alongside changes in music production and music consumption. While previous studies have shown that Spotify audio features are effective for predicting song popularity [2], [3], [4], the present

study extends this line of research by demonstrating that the relative importance of these features changes systematically across historical periods. Rather than representing a fixed set of predictors, the acoustic profile of successful songs adapts to technological and cultural changes within the music industry.

Among the identified trends, the increasing importance of instrumentality and loudness, together with the declining influence of speechiness and acousticness, suggests a gradual shift towards increasingly vocal-oriented and highly produced recordings. These patterns are consistent with long-term developments in commercial music production and broadly agree with previous observations that successful contemporary songs tend to exhibit greater loudness and lower instrumental content [4], [9]. The statistical significance tests presented in Section 4.3, together with the robustness analyses in Section 4.4, provide converging evidence that these temporal patterns are unlikely to be explained by sampling variability or model specification alone. Although the present study does not establish causal relationships, the consistent temporal patterns observed across multiple robustness analyses indicate that these changes reflect stable characteristics of the data rather than modelling artifacts.

Importantly, these findings should be interpreted within the scope of the predictor variables analysed. The proposed models explain how intrinsic audio characteristics contribute to distinguishing hit songs from flops, rather than providing a comprehensive explanation of commercial success itself. Real-world chart performance is also influenced by numerous external factors, including artist popularity, marketing activities, playlist placement, release timing, social media exposure, audience engagement, and record-label support, which were intentionally excluded from the present study because they are not represented in the Spotify audio-feature dataset. Consequently, the reported feature-importance patterns should be interpreted as describing the evolving contribution of musical attributes to commercial success, rather than the complete set of determinants underlying hit-song popularity.

Beyond the music domain, this study demonstrates the value of combining machine learning with Explainable Artificial Intelligence for analyzing temporal evolution in complex datasets. Recent research has demonstrated that changes in SHAP-based feature importance can provide interpretable evidence of temporal changes in model behaviour and data characteristics [24], [25]. By comparing SHAP-derived feature importance across independently trained era-specific models and validating the results through multiple robustness analyses, the proposed framework provides a practical approach for investigating how predictive relationships change over time. This analytical perspective may also be applicable to other longitudinal prediction problems in applied data science where both predictive performance and model interpretability are of interest.

5. Conclusion

This study investigated the temporal evolution of hit-song predictors using Spotify audio features extracted from Billboard-labelled tracks released between 1960 and 2019. Era-specific machine learning models were developed to compare predictive performance across three music-industry eras, while SHAP was employed to examine how the importance of audio features evolved over time. Statistical bootstrap testing was further used to determine whether the observed cross-era changes in feature importance exceeded sampling variability, while multiple robustness and sensitivity analyses evaluated the stability of the findings. Among the evaluated classifiers, Random Forest consistently achieved the strongest predictive performance across all eras. More importantly, the explainability analysis demonstrated that the determinants of commercial success are not temporally stationary. Instrumentality remained the most influential predictor throughout the study period, loudness became increasingly important in the Streaming era, whereas the influence of speechiness and acousticness gradually declined.

The principal changes in feature importance were supported by bootstrap confidence intervals and statistical significance testing, and remained consistent across repeated model fitting, alternative era definitions, data de-duplication, and model-agnostic feature-importance analysis, providing robust evidence that the audio characteristics associated with commercially successful songs have evolved over time. The reported findings should be interpreted as describing the evolving contribution of intrinsic audio characteristics rather than the complete set of determinants underlying commercial success, since external influences such as artist popularity, marketing activities, playlist placement, and audience engagement were beyond the scope of the available dataset.

Overall, this study demonstrates that combining machine learning with Explainable Artificial Intelligence provides an effective framework for investigating temporal changes in predictive relationships beyond conventional model

evaluation. The proposed approach offers a practical methodology for understanding the long-term evolution of complex cultural datasets and may be applicable to other domains where predictor importance changes over time. Future research could extend this framework by incorporating lyrical, artist, and listener-related information, as well as applying formal concept-drift detection techniques to identify more precise transitions in hit-song characteristics.

6. Declarations

6.1. Author Contributions

Conceptualization: C.A. and F.S.; Methodology: C.A.; Software: F.S.; Validation: C.A. and F.S.; Formal Analysis: C.A. and F.S.; Investigation: F.S.; Resources: C.A.; Data Curation: F.S.; Writing Original Draft Preparation: C.A. and F.S.; Writing Review and Editing: C.A. and F.S.; Visualization: F.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors gratefully acknowledge the financial support provided by the School of Information Technology, Universitas Ciputra, for this research.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] IFPI, "Global music report 2025: Global recorded music revenues grew 4.8% in 2024," 2025. [Online]. Available: <https://www.ifpi.org/ifpi-amidst-highly-competitive-market-global-recorded-music-revenues-grew-4-8-in-2024/>
- [2] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, and N. L. Komarova, "Musical trends and predictability of success in contemporary songs in and out of the top charts," *R. Soc. Open Sci.*, vol. 5, no. 5, pp. 1–10, May 2018, doi: 10.1098/rsos.171274.
- [3] D. Herremans, D. Martens, and K. Sørensen, "Dance hit song prediction," *J. New Music Res.*, vol. 43, no. 3, pp. 291–302, 2014, doi: 10.1080/09298215.2014.881888.
- [4] H. S. Saragih, "Predicting song popularity based on Spotify's audio features: Insights from the Indonesian streaming users," *J. Manage. Anal.*, vol. 10, no. 4, pp. 693–709, 2023, doi: 10.1080/23270012.2023.2239824.
- [5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, no. 1, pp. 4765–4774, Dec. 2017.
- [6] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
- [7] X. Guo, "The evolution of the music industry in the digital age: From records to streaming," *J. Sociol. Ethnol.*, vol. 5, no. 10, pp. 7–12, 2023, doi: 10.23977/jsoc.2023.051002.
- [8] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, and N. L. Komarova, "Musical trends and predictability of success in contemporary songs in and out of the top charts," *R. Soc. Open Sci.*, vol. 5, no. 5, pp. 1–10, May 2018, doi: 10.1098/rsos.171274.

-
- [9] J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos, “Measuring the evolution of contemporary Western popular music,” *Sci. Rep.*, vol. 2, no. 1, pp. 1–8, 2012, doi: 10.1038/srep00521.
- [10] D. Byrne, *How Music Works*. San Francisco, CA, USA: McSweeney’s, 2012.
- [11] M. A. Morris and D. Powers, “Control, curation and musical experience in streaming music services,” *Creative Ind. J.*, vol. 8, no. 2, pp. 106–122, 2015, doi: 10.1080/17510694.2015.1090222.
- [12] Y. Dou, J. Cui, and X. Wu, “Comparative analysis of multiple machine learning algorithms in heart disease classification prediction,” *IEEE Int. Conf. Inf. Technol. Mechatronics Eng.*, vol. 2025, no. Mar., pp. 1212–1218, Mar. 2025, doi: 10.1109/ITOEC63606.2025.10967616.
- [13] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, “Explanation of machine learning models using Shapley additive explanation and application for real data in hospital,” *Comput. Methods Programs Biomed.*, vol. 214, no. 1, pp. 1–12, 2022, doi: 10.1016/j.cmpb.2021.106584.
- [14] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [15] N. Askin and M. Mauskopf, “What makes popular culture popular? Product features and optimal differentiation in music,” *Am. Sociol. Rev.*, vol. 82, no. 5, pp. 910–944, 2017, doi: 10.1177/0003122417728662.
- [16] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
- [17] R. Couronné, P. Probst, and A.-L. Boulesteix, “Random forest versus logistic regression: A large-scale benchmark experiment,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, 2018, doi: 10.1186/s12859-018-2264-5.
- [18] S. M. Lundberg, G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” *arXiv*, vol. 2018, no. Feb., pp. 1–10, Feb. 2018.
- [19] J. Yang, “Fast TreeSHAP: Accelerating SHAP value computation for trees,” *arXiv*, vol. 2021, no. Sep., pp. 1–10, Sep. 2021.
- [20] The Overman, “The Spotify hit predictor dataset (1960–2019),” Kaggle, 2019. [Online]. Available: <https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset>
- [21] N. V. D. S. S. V. P. Raju and P. N. Devi, “A comparative analysis of machine learning algorithms for big data applications in predictive analytics,” *Int. J. Sci. Res. Manage.*, vol. 12, no. 10, pp. 896–907, Oct. 2024, doi: 10.18535/ijsrc/v12i10.ec09.
- [22] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, no. 1, pp. 1–12, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.
- [23] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2825–2830, 2011.
- [24] C. Duckworth, F. P. Chmiel, D. K. Burns, et al., “Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-02481-y.
- [25] R. Çınar, G. Yuksek, T. Lale, S. Ekinici, D. Izci, and A. Ma’arif, “SHAP-based framework for temporal detection of sensor drift in gas sensor arrays,” *J. Robot. Control*, vol. 6, no. 5, pp. 1–10, 2025, doi: 10.18196/jrc.v6i5.28048.