“VIRAL VISION”
CLASSIFICATION MODEL
“SONG POPULARITY”

EXECUTIVE SUMMARY

In the competitive world of music, identifying potential hits is crucial for survival. While the number of successful songs is soaring, exceeding industry revenue growth, the earnings potential per song is shrinking. This creates a critical need for data-driven tools that navigate the saturated market and optimize resource allocation. This report introduces ViralVision's innovative song success predictor, an ML-powered solution designed to empower record labels with actionable insights, maximizing ROI and minimizing risk. We are confident that our tool can become a valuable asset in your journey to discover the next chart-topping anthem.

TODAY'S MARKET

After weathering some tough years, the music industry is showing signs of life with steady growth over the past eight years. However, a major shift from physical formats to streaming platforms has fundamentally reshaped how people consume music. While this digital transformation poses challenges like lower per-stream revenue and fierce competition, it also unlocks exciting opportunities. Streaming platforms offer artists wider reach and valuable data insights for development and marketing. Moving forward, record labels that adapt to the digital landscape and embrace innovation will be best positioned to thrive in this exciting new era of music.

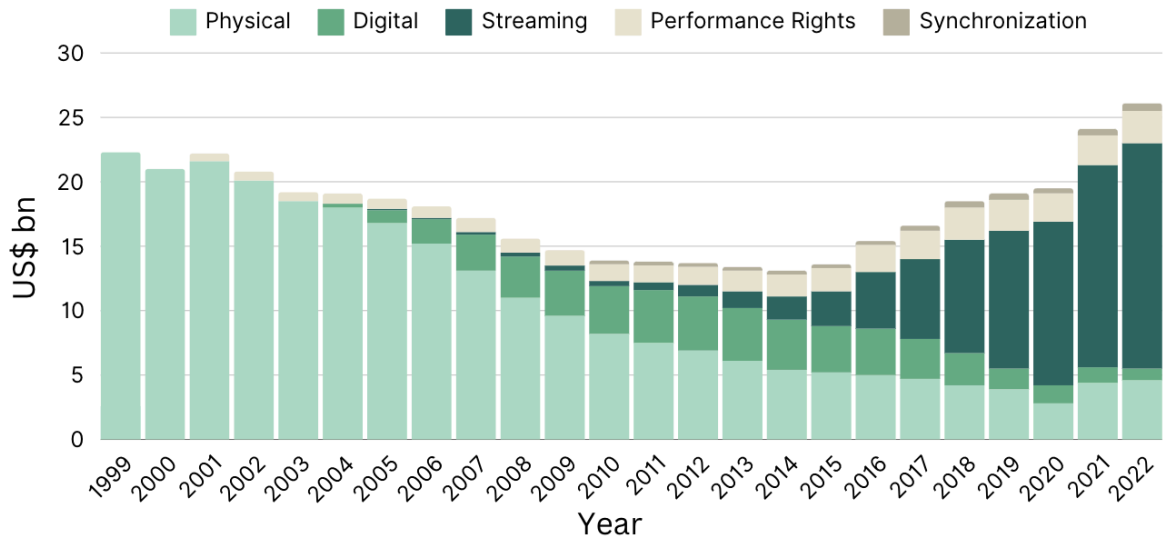


Chart displaying revenue in the music industry over the years with a breakdown by medium

In recent times, the prevalence of successful songs has surged to unprecedented levels, with this trend showing no signs of slowing down. Despite this surge in hit songs, revenue growth

within the sector has been comparatively sluggish. This indicates that while achieving success may be more attainable, the financial returns are diminishing. Consequently, there's a pressing need to optimise the revenue generated by successful songs.

In light of this, it's more important than ever to ensure that investments are directed towards the right songs, especially given the downward trend in revenue. With resources becoming scarce, making informed decisions about where to allocate funds is paramount to maximising profitability and sustaining growth.

CASE STUDY - UNEXPECTED FAILURE

An illustrative case of an album release failing to meet its expected return on investment (ROI) due to inaccurate prediction is exemplified by "Chinese Democracy" by Guns N' Roses.

"Chinese Democracy" was highly anticipated, marking the return of Guns N' Roses boasting a reported production cost of approximately $13 million, expectations were sky-high. However, despite the considerable investment and anticipation, the album received mixed reviews and fell short of the commercial success achieved by the band's previous works. This instance underscores the importance of accurate prediction in album/song releases to avoid such disappointments. **ViralVision predicted that the song would not be a success.**

VIRALVISION

THE PRODUCT

At its core, our predictor harnesses the power of data to provide valuable insights into the potential success of a song. Operating in simple, non-technical terms, it analyses a multitude of factors including artist popularity, musical features, and genre to compute the probability of a song's success. Think of it as a crystal ball for the music industry, offering predictive capabilities that were once unimaginable.

But what sets our predictor apart are its key features and functionalities. Utilising the latest advancements in machine learning and data analytics, our algorithm has been meticulously trained on the most comprehensive dataset available. Spanning the entirety of the 21st century, this dataset comprises detailed information on the most successful songs since 2000, providing a rich tapestry of insights into what makes a hit.

By going deep into this wealth of data, our predictor identifies the underlying patterns and trends that separate chart-toppers from the rest. It's not just about analyzing individual songs; it's about understanding the broader landscape of the music industry and deciphering what resonates most with audiences.

From the infectious beats of pop to the soul-stirring melodies of R&B, our predictor is equipped to handle a diverse range of genres. It doesn't just stop at genre classification; it dives deeper, examining the nuances and subtleties that define each musical style. By doing so, it ensures that its predictions are finely tuned to the unique characteristics of each song.

In essence, our song success predictor represents a paradigm shift in the way we approach music. It's more than just a tool; it's a glimpse into the future of the industry, where data-driven insights pave the way for unprecedented success.

PRODUCT TECHNICAL APPROACH

ViralVision leverages the cutting-edge CatBoost algorithm, deliberately chosen for its unparalleled reliability and precision, representing the forefront of modern algorithms. CatBoost, functioning as a tree algorithm, iteratively constructs multiple trees, with each iteration refining its predictions based on the errors of the preceding ones, resulting in an exceptionally robust predictive model. Through feature importance analysis, we gain insights into the critical factors that influence our model, enabling us to understand in detail the factors contributing to a song's success, surpassing mere prediction of success or failure. Our model achieves an accuracy of approximately 70% or higher, depending on the desired level of predictive success.

2023 ADVISE CASE-STUDY

By leveraging feature importance analysis, we can generate predictions for an ideal song with the greatest potential for success. In line with this, our model for the year 2023 suggested releasing a song characterized by high acousticness, low speechiness, shorter duration, minimal liveness, and low danceability. "Te Felicito" by Shakira & Rauw Alejandro closely adhered to these recommendations, as suggested by ViralVision.

Now, with the insights gained in 2024, we can confirm that our model accurately predicted a successful song, as evidenced by its prominent placement on charts worldwide. This validation underscores the effectiveness and reliability of our predictive algorithms in identifying hit songs.

PRODUCTIVITY BENEFITS

The implementation of ViralVision offers multifaceted benefits to record labels, encompassing various aspects of their operations. Firstly, the tool significantly enhances the return on investment (ROI) by 15-35% for marketing and promotional activities. By accurately predicting the success of songs, ViralVision enables labels to allocate resources more efficiently. This means that marketing efforts can be strategically focused on tracks with the highest potential for success, thereby maximising the ROI on promotional campaigns.

ARTIST DEVELOPMENT BENEFITS

Moreover, ViralVision plays a pivotal role in strategic artist development. It goes beyond evaluating individual songs and provides valuable insights into artists themselves. This allows record labels to make informed decisions regarding artist signings and development initiatives. By directing attention towards artists and music styles with the highest likelihood of success in the current market landscape, ViralVision empowers labels to nurture promising talent effectively.

In addition to its impact on marketing and artist development, ViralVision also facilitates enhanced production decisions. Leveraging insights from the tool, labels can make informed choices regarding various aspects of production, including genre, tempo, and lyrical themes. By aligning these decisions with current trends and audience preferences, labels can create music that resonates more deeply with their target demographic, thereby increasing the likelihood of commercial success.

RISK MANAGEMENT

ViralVision serves as a crucial tool for risk mitigation within the music industry. The unpredictable nature of the market often poses significant financial and reputational risks for record labels, particularly when investing heavily in new songs or artists. However, by providing data-driven insights into the potential success of songs and artists, ViralVision enables labels to make smarter, less risky decisions. This ultimately helps to safeguard their financial resources and protect their reputation within the industry.

CATALOGUE OPTIMIZATION

Finally, ViralVision offers opportunities for catalogue optimization, not only focusing on new releases but also analysing existing catalogues. By identifying potential sleeper hits within their catalogue, labels can unlock opportunities for re-promotion or repackaging. This allows them to reach new audiences and maximise revenue potential from their existing repertoire of music.
In implementing ViralVision within your record label, we outline a streamlined process designed to seamlessly integrate this cutting-edge solution into your existing operations while addressing any potential concerns.

CONCLUSION

In closing, we invite you to seize the opportunity to revolutionise your record label's approach to music industry success with our song success predictor. By harnessing the power of data-driven insights, ViralVision offers unparalleled value in optimising marketing ROI, strategically developing artists, enhancing production decisions, mitigating risks, and optimising catalogue performance.

TECHNICAL ANNEX

DATA PREPROCESSING

FEATURE CREATION

After performing data exploration, firstly null values were handled. As only one data observation contains null values, removing it from the dataset was the decision taken.
Non-relevant features for the model, 'Track Name', 'Album' and 'Artist Name' were removed due to irrelevance to track success, as we already had artist popularity. As the 'Key' feature is a categorical variable, it was transformed into a binary table.

As an additional feature, a different model perspective was created for business models in which album popularity is known, e.g. acquiring specific music rights, or investing into artists albums. As this feature doesn't represent all business models the model will be tested twice with and without this feature. Furthermore, the average of popularity per genre was also added as a feature.

Finally, in order to achieve better explainability of the models regarding feature importance, a random feature ('*random_feature*') was included as the threshold that separates the relevant features from irrelevant features.

APPROACH TO ARTIST GENRES

Each track contains the artist's genres. However, since artists usually produce music across various genres, amounting to 517 different "subgenres," there was a need to determine the main genre represented by the artist and linking it to the song. To achieve this, a mapping was created to associate each subgenre with its corresponding main genre. Then, Naive Bayes, a multiclass classifier, was employed to predict the most probable main genre for all tracks. **Table 2** - *Main Genre Identification with Naive Bayes multi-classifier.* (1 - With Naive Bayes, 2- without Naive Bayes).

Through this method, dimension reduction, noise reduction, and an improved sample size were achieved, along with overall simplification of the data (without sacrificing relevant information). This enhances accuracy, precision, recall, F1-score, and most importantly, reduces the risk of overfitting. **Table 3** - *Evaluation metrics for the usage of Naive Bayes multi-classifier*

While simply measuring accuracy doesn't show much difference with adding Naive Bayes, it's crucial to remember that accuracy can be misleading, especially when dealing with imbalanced data like song success (hits vs. misses). That's why metrics like F1-score and AUC, which are more robust in such cases, reveal subtle but important improvements in Bagging and Random Forest models. Even small shifts in these metrics can translate to a significant difference in predicting a successful track.

After feature engineering the dataset used for the model had 2229 rows and 36 features (columns). **Table 1** - *Dataset after Feature Engineering example.*

CREATING POPULARITY BINS

The implementation of this process brought empirical evidence that the most influential process in model's performance was the way popularity bins were designed - thresholds delimiting from which value of popularity a song becomes successful. A theme that can be extremely subjective. Four strategies were experimented: Mean Binning, 3-Quantiles, 4-Quantiles and Decision Tree Binning. **Table 4** - *Model Accuracy for different binning strategies*.

The different performances in the different strategies can be explained by the dataset available. The dataset only rates the popularity of the top popular 100 songs per year, which is already an indicator of popularity itself. Therefore, as expected, the popularity is skewed to the right. The outliers present in the left (less popular songs), can be explained by the generational effect on which a popular song in previous years might not be as popular if launched in the present. **Picture 1** - *Popularity feature Distribution.*

Moreover, the dataset only contains 2229 observations, which when binning popularity using 4-quantile and 3-quantile, might not be enough, (approx. 445 and 595 training data, respectively) explaining the low accuracy in these binning strategies.

Decision tree binning discovered data-driven popularity classes, outperforming fixed binning methods in predicting song success, captured non-linear relationships between popularity and success, and improved model accuracy. This might happen because unlike Mean Binning's equal halves, Decision Tree's data-driven thresholds captured popularity clusters & non-linearity, leading to superior prediction of song success in our skewed dataset.

While Decision Tree Binning achieved the highest accuracy in predicting song success, it's crucial to remember that model performance is ultimately a trade-off with business requirements.

APPROACH TO CLASSIFICATION ANALYSIS

After binning the data the approach was performing different classification methods in order to extract the one with the best performance. Regression analysis could also be applied to the problem as the success threshold is still subjective to choose, but then in terms of marketability of the product it would not add much value. Therefore, the followed approach was classification to a binary classification, Successful and Not-Successful.

MODELS AND METHODS USED

To find the fittest model to this dataset, we used a wide range of models. Starting with a single Decision Tree, with enhanced explainability the features who provided the best decrease of impurity in the data. Followed by Bagging and Random Forest which capture the diversity of the data through ensemble learning, reducing variance and potentially improving overall accuracy (**Table 5** - *Model Performance across models*) Random Forest further refines this approach by introducing randomness in feature selection at each split, mitigating overfitting and enhancing generalizability.

Moving beyond individual trees, we explored K-Nearest Neighbors (KNN) for its simplicity and ability to capture complex, non-linear relationships. Unlike Decision Trees with their predefined rules, KNN predicts based on the majority vote of its nearest neighbours in the training data. This approach thrives on continuous features and intricate patterns, potentially offering insights beyond Decision Trees.

Venturing further, we encountered the powerful CatBoost algorithm, renowned for its gradient boosting prowess and flexibility. Unlike Decision Trees, CatBoost sequentially builds upon decision trees, correcting errors from previous iterations. Like the other models, we meticulously tuned CatBoost's hyperparameters through grid search, ensuring it reached its full potential in our specific context. **Table 6** - *Best Parameters per Model*

TRADE-OFF IN MODEL EVALUATION

The heatmap (**Picture 2** - *Model Performance Metrics*) provides a comparison of predictive models, each evaluated on key performance metrics, with the goal of identifying successful tracks. Catboost emerges as a robust choice, boasting the highest scores in F1 and ROC AUC metrics. These scores indicate a strong balance between precision (minimising false positives) and recall (capturing true hits), as well as an excellent ability to discriminate between successful and unsuccessful tracks across all thresholds.

While other models like RFGrid and DTGrid excel in individual metrics such as accuracy and precision, Catboost's consistent performance across all measures suggests it's less prone to the trade-offs inherent in model selection. In practice, this means Catboost is likely to provide a reliable identification of hits while minimising investment in likely non-successful tracks, an essential aspect when predicting track success.

The subpar Recall, F1-score, and Precision observed in the heatmap can be traced back to specific limitations within the dataset. One primary issue is the dataset's biassed nature: it's not evenly distributed across different music genres, with some genres being underrepresented. This imbalance likely hampers the models' ability to generalise, which is critical for accurately identifying successful songs across a diverse range of styles. Furthermore, the dataset is relatively small, comprising just 2229 observations—a size that's insufficient for the complex models in question, which require more extensive data to detect subtle patterns and make reliable predictions.

Another significant limitation is the dataset's current popularity metrics, which do not consider the temporal aspects of song popularity. The absence of this temporal data prevents the models from capturing trends and patterns over time that could be instrumental in predicting a track's success. Moreover, the dataset's focus on only the top 100 popular songs per year creates a narrow definition of success. This restriction likely excludes many tracks that could provide valuable insights into the broader characteristics of successful music, thus constraining the models' understanding and leading to poorer performance in key metrics.

THE TOP OF THE CLASS MODEL

After taking into account all performance metrics and business requirements, our top of the class model is CatBoost with optimised hyperparameters (Best Parameters for CatBoost - Table 6 - *Best Parameters per Model*).

CatBoost demonstrates a commendable balance between precision and recall, as indicated by its leading F1 and ROC AUC scores. This balance is crucial for our goal of accurately predicting successful songs, minimising false positives while effectively identifying true hits. CatBoost's ability to maintain consistent performance across various measures, despite the dataset's limitations, solidifies its position as the most reliable model for our needs in the music industry.

APPROACH TO ARTIST AND PRODUCERS RECOMMENDATIONS

FEATURE IMPORTANCE

Aiming to create an accurate artist recommendation system, we employed a unique approach to identify the most relevant features for predicting user preferences. This methodology leverages the concept of a "random feature," which serves as a control or baseline comparison. By analysing the feature only above the random feature, we gain valuable insights into which features truly contribute to accurate recommendations. (**Picture 3 - *Feature Importance with Random Feature***).

FEATURE INFLUENCE

For more valuable artist recommendations we enter the SHAP summary plot, our decoder ring for understanding individual feature influence. (**Picture 4 - *Individual Feature Influence***). We used the features of Random Forest, because Catboost doesn't allow this analysis, and Random Forests is the second best performer.

Think of it as a ranked chart, with top features boasting the most sway over predicted popularity. But it's not just about rank - colour matters! Reds ignite popularity, blues dampen it, and intensity reflects impact strength. Each coloured area showcases the distribution of feature values, revealing where the model has "seen" specific features in action. Don't miss

the horizontal lines – baselines representing the average impact of each feature on popularity. Steeper lines signal stronger influencers.

This plot unlocks a treasure trove: analysing nuances of individual feature impact, and uncovering potential interactions where features collaborate or clash.

TABLE ANNEX

INDEX	year	artist_popularity	danceability	energy	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	avg_album_popularity
0	2000	86	0.429	0.661	85.0	1.0	0.0281	0.00239	0.000121	0.2340	0.285	173.372	266773.0	91.00
1	2000	75	0.434	0.897	69.0	1.0	0.0488	0.01030	0.000000	0.6120	0.684	148.726	167067.0	84.00
2	2000	61	0.529	0.496	95.0	1.0	0.0290	0.17300	0.000000	0.2510	0.278	136.859	250547.0	64.50
3	2000	83	0.556	0.864	76.0	0.0	0.0584	0.00958	0.000000	0.2090	0.400	105.143	216880.0	88.00
4	2000	65	0.610	0.926	68.0	0.0	0.0479	0.03100	0.001200	0.0821	0.861	172.638	200400.0	70.25
5	2000	56	0.706	0.888	84.0	1.0	0.0654	0.11900	0.000096	0.0700	0.714	121.549	253733.0	73.00
6	2000	88	0.949	0.661	62.0	0.0	0.0572	0.03020	0.000000	0.0454	0.760	104.504	284200.0	86.00
7	2000	69	0.712	0.762	63.0	1.0	0.0326	0.02600	0.000000	0.0981	0.842	103.032	260560.0	57.00
8	2000	69	0.713	0.678	54.0	0.0	0.1020	0.27300	0.000000	0.1490	0.734	138.009	271333.0	76.50
9	2000	80	0.458	0.795	51.0	1.0	0.0574	0.00316	0.000202	0.0756	0.513	123.229	255373.0	83.00

INDEX	radom_feature	avg_genre_popularity	main_genre_Alternative	main_genre_Electronic	...	key_0.0	key_1.0	...2	popularity
0	0.093456	71.405797	0	0	...	0	0	...	1
1	0.592555	72.768939	0	0	...	1	0	...	1
2	0.786159	64.683673	1	0	...	0	0	...	0
3	0.958255	72.768939	0	0	...	0	0	...	1
4	0.870356	71.405797	0	0	...	0	0	...	0
5	0.905640	70.681319	0	0	...	0	0	...	0
6	0.195571	70.681319	0	0	...	0	0	...	1
7	0.157867	72.768939	0	0	...	0	0	...	0
8	0.793675	71.405797	0	0	...	0	0	...	1
9	0.209057	72.768939	0	0	...	1	0	...	1

Table 1 - *Dataset after Feature Engineering example*

INDEX	main_genre	artist_genres
0	Pop	['permanent wave'; 'pop']
1	Rock	['alternative metal'; 'modern rock'; 'pop punk'; 'punk'; 'rock'; 'socal pop punk']
2	Alternative	['contemporary country'; 'country'; 'country dawn'; 'country road']
3	Rock	['alternative metal'; 'nu metal'; 'post-grunge'; 'rap metal'; 'rock']
4	Pop	['boy band'; 'dance pop'; 'pop']
5	Hip Hop	['contemporary r&b'; 'dirty south rap'; 'hip pop'; 'r&b'; 'urban contemporary']

Table 2 - *Main Genre Identification with Naive Bayes multi-classifier*

Model	Accuracy 1	Accuracy 2	Precision 1	Precision 2	Recall 1	Recall 2	F1-Score 1	F1-Score 2	AUC 1	AUC 2
Decision Tree	0,9413	0,9413	0,9283	0,9283	**0,8644**	0,8644	0,8831	0,8831	0,9161	**0,9161**
Bagging	93,86% (94,57%)	0,9353	0,9189	0,9478	**0,8644**	0,9196	0,8908	0,8957	0,9109	**92.33%**
Random Forest	93,58% (94,35%)	94,35%(93,47%)	0,9107	0,9035	**0,8644**	0,8729	0,887	0,8879	0,9176	**0,9204**
Boosting (XGBoost)	0,9413	0,9413	0,9413	0,9413	**0,9413**	0,9413	0,9413	0,9413	0,9413	**0,9413**
KNN	0,8022	0,7783	0,6957	0,5952	**0,4069**	0,4237	0,5134	0,495	0,6727	**0,6622**
CatBoost	0,9457	0,9478	0,9115	0,9196	**0,8729**	0,8729	0,8918	0,8957	0,9218	**0,9233**

Table 3 - Evaluation metrics for the usage of Naive Bayes *multi-classifier*

Model	4-Quantile	Decision Tree	3-Quantile	Mean
Decision Tree	39,35%	**67,39%**	50,00%	**67,83%**
Bagging	46,09%	**78,14%**	55,65%	**72,92%**
Random Forest	46,52%	**77,54%**	55,35%	**73,30%**
Boosting (XGBoost)	-	**77,83%**	-	**70,43%**
KNN	40,35%	**74,55%**	51,28%	**69,17%**
CatBoost	-	**69.78%**	-	**69,78%**

Table 4 - *Model Accuracy for different binning strategies*

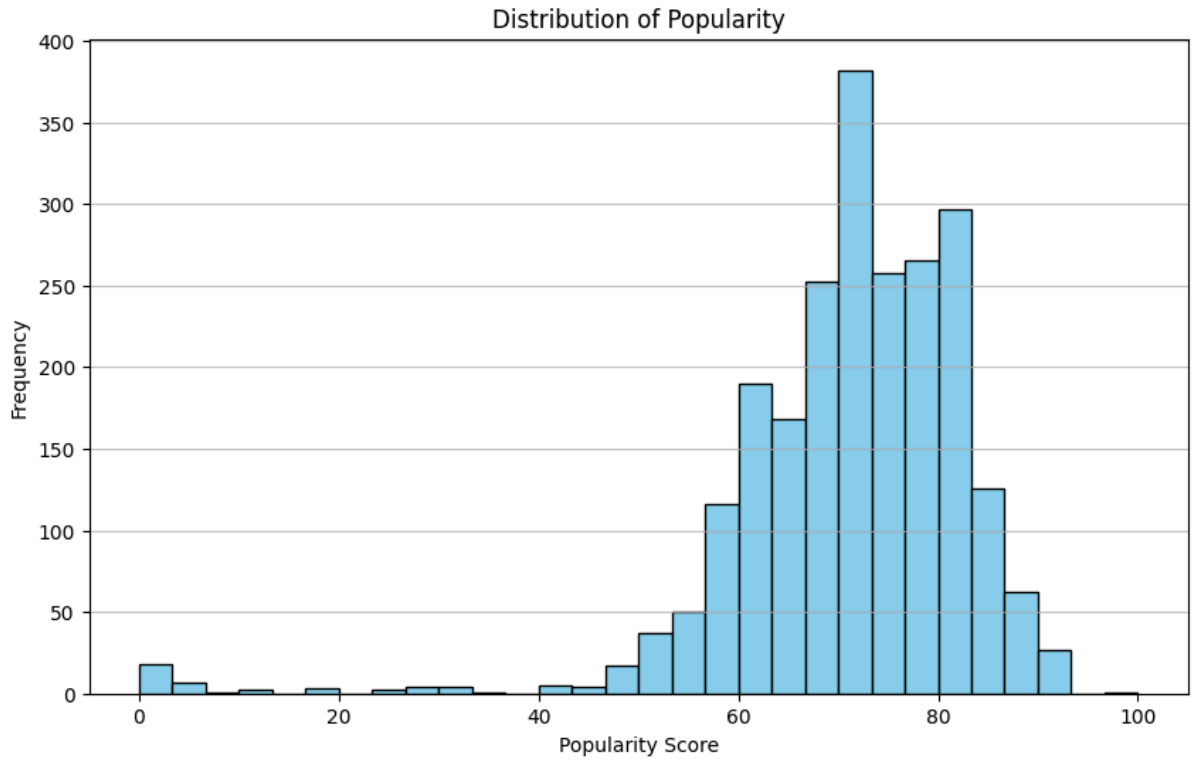
Model	Accuracy	Precision	Recall	F1-Score	AUC
Decision Tree	67,39%	**62,12%**	54,24%	**47,23%**	64,11%
Bagging	78,14%	**57,81%**	31,36%	**40,66%**	61,73%
Random Forest	77,54%	**53,41%**	39,83%	**45,63%**	63,92%
Boosting (XGBoost)	77,83%	**56,00%**	35,59%	**43,52%**	62,97%
KNN	74,55%	**43,48%**	25,42%	**32,09%**	57,01%
CatBoost	69.78%	**44,38%**	66,95%	**53,38%**	69,00%

Table 5 - *Model Performance across models*

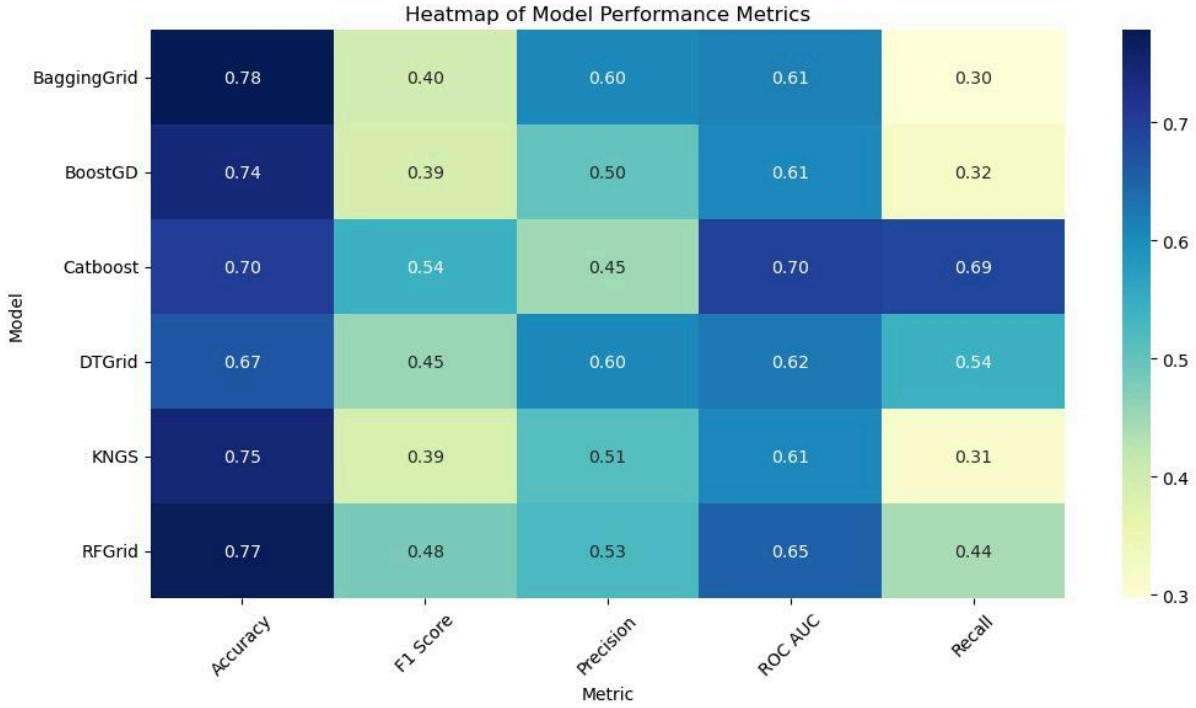
Model	Best Parameters
Decision Tree	max_depth: 14
Bagging	n_estimators: 300
Random Forest	max_features: 25, min_samples_leaf: 3
Boosting (CatBoost)	learning_rate: 0.1, max_depth: 7, n_estimators: 300, subsample: 1.0
K-Nearest Neighbors	n_neighbors: 11, metric: 'manhattan', weights: 'distance'
Gradient Boosting (XGBoost)	depth: 8, learning_rate: 0.01, n_estimators: 300, subsample: 0.6

Table 6 - Best Parameters per Model

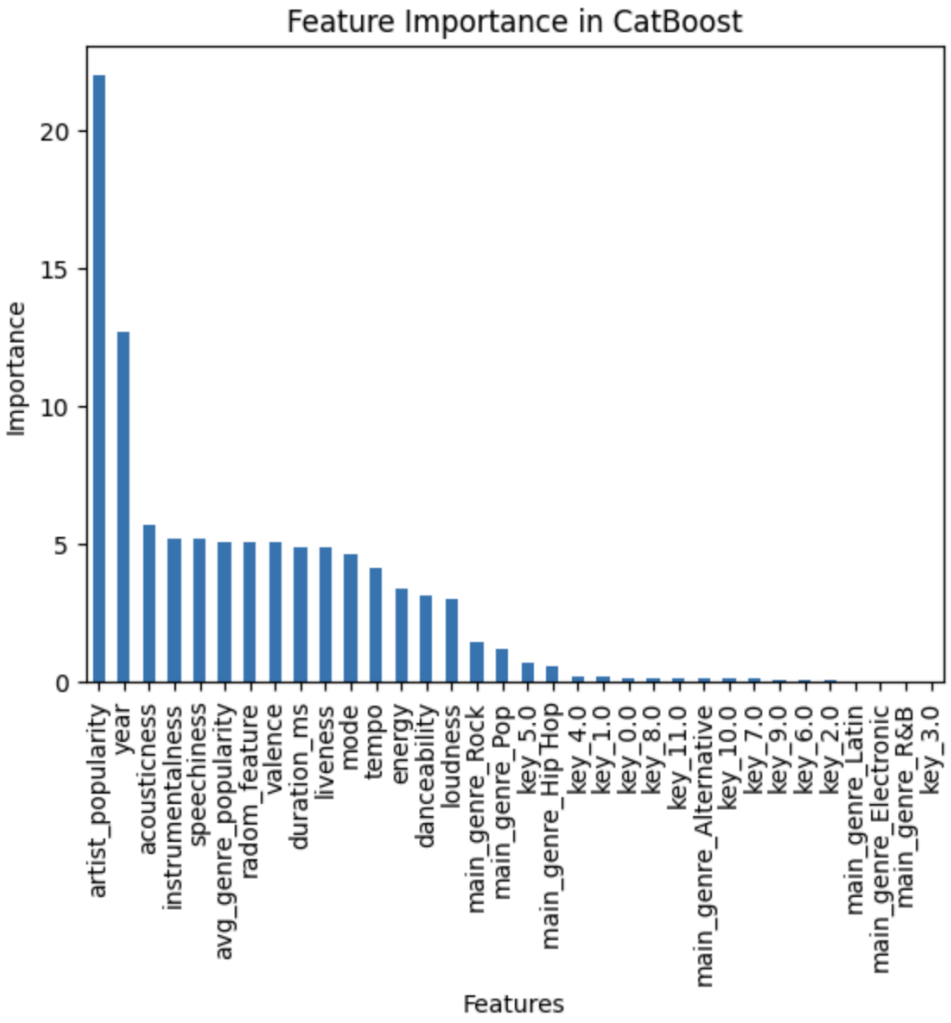
Picture Annex



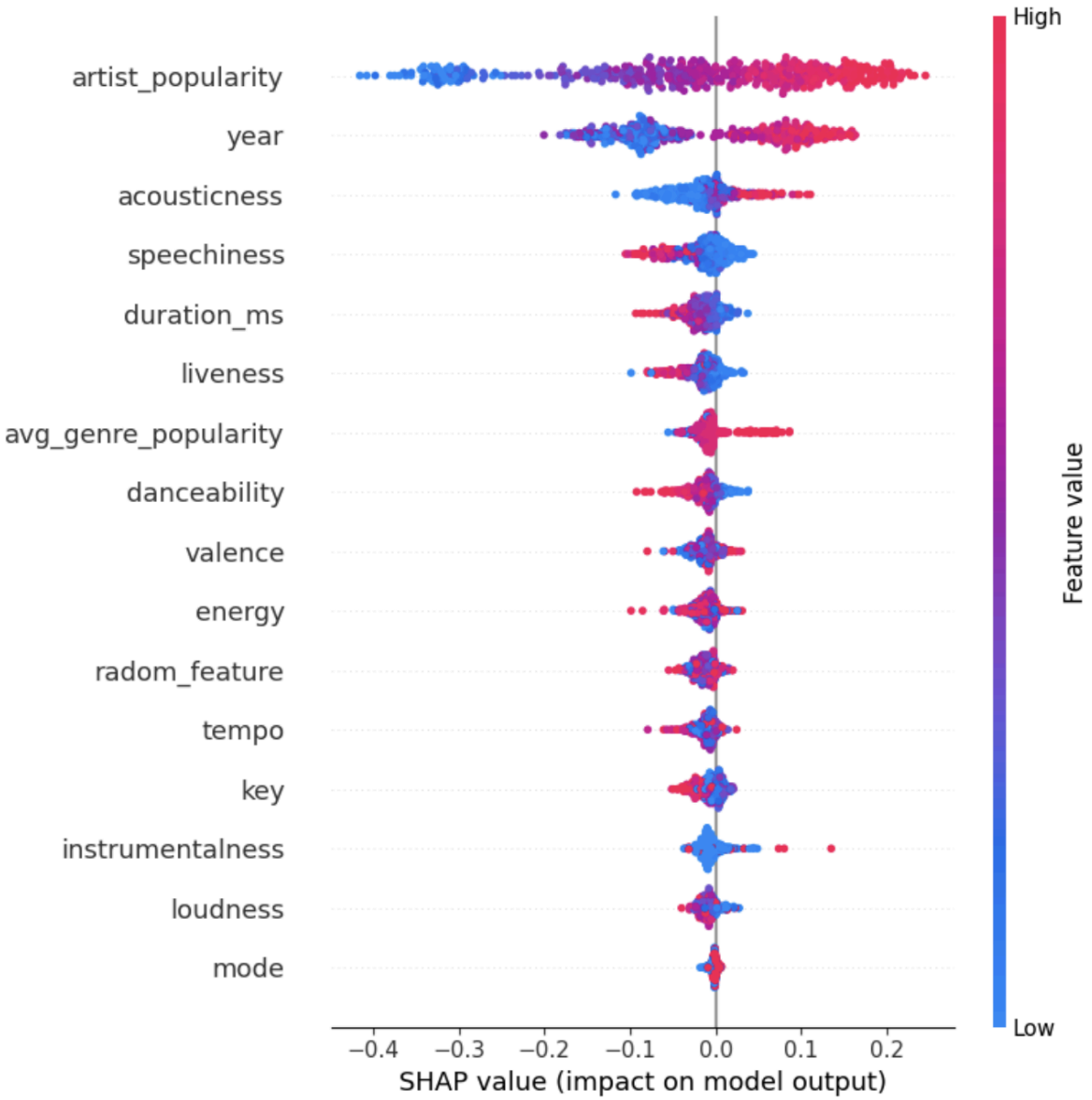
Picture 1 - *Popularity feature Distribution.*



Picture 2 - Model Performance Metrics



Picture 3 - *Feature Importance with Random Feature*



Picture 4 - *Individual Feature Influence*