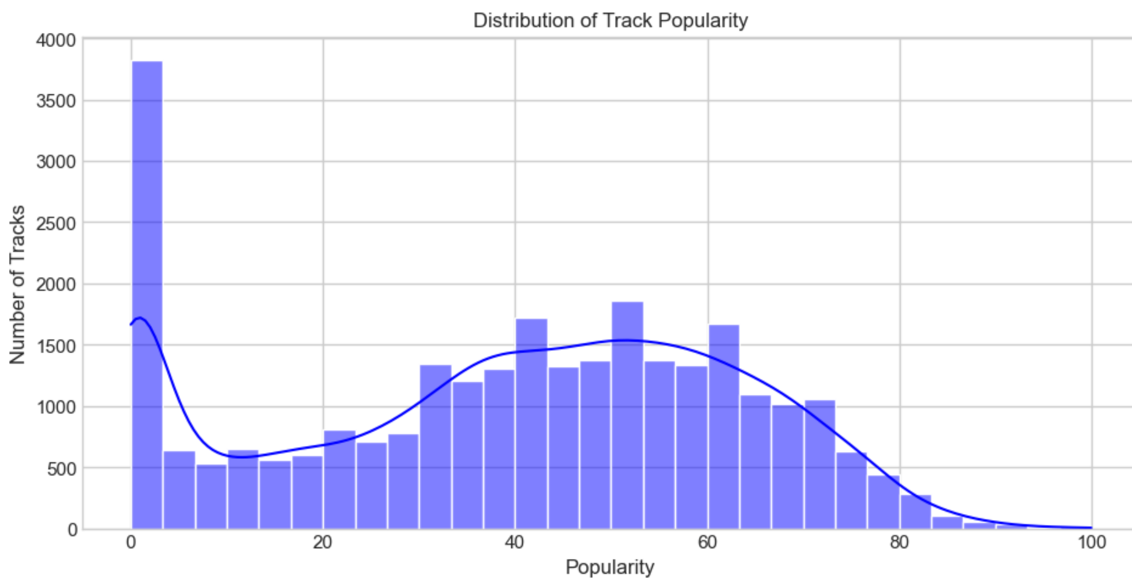
Spotify Track Popularity Prediction – Visual Report

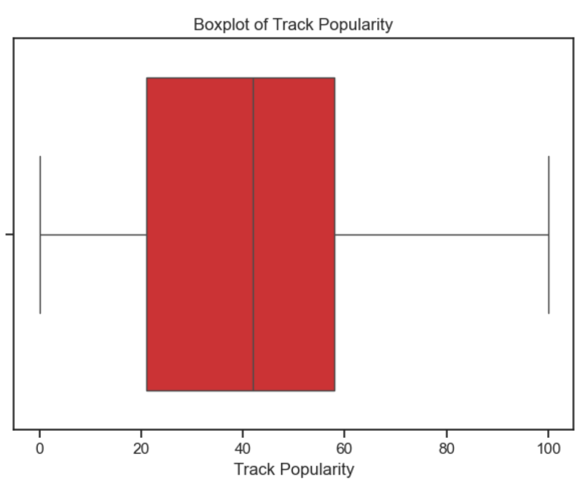
1. Distribution of Track Popularity

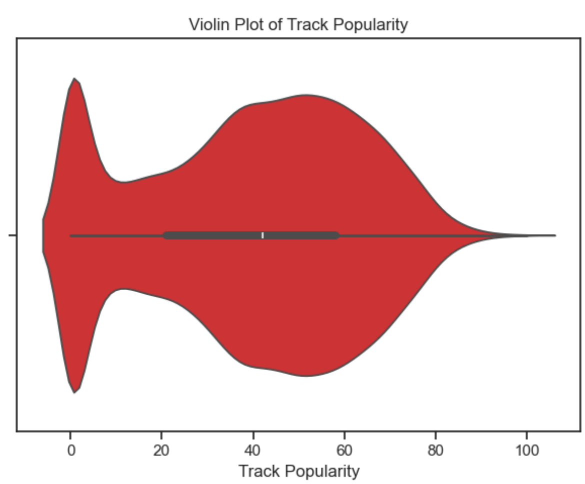
This histogram with KDE shows that most tracks have low to moderate popularity, while fewer songs reach high popularity levels. The distribution is clearly right-skewed.



2. Boxplot of Track Popularity

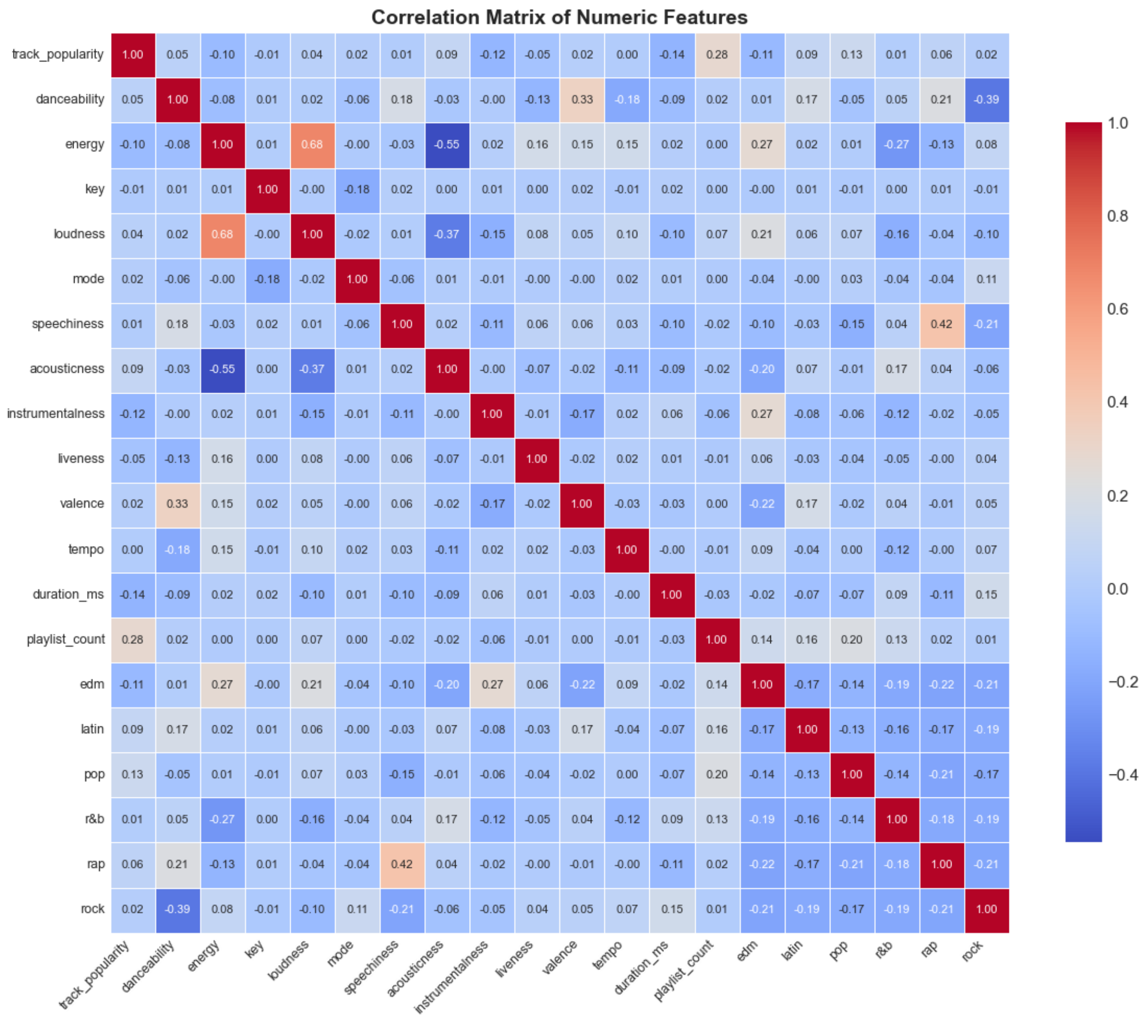
The boxplot highlights outliers in track popularity. Some songs reach extremely high scores, which can influence model training if not treated.





3. Correlation

This correlation heatmap reveals strong relationships between certain audio features, such as energy and loudness. It helps in identifying redundant variables and guides feature selection.

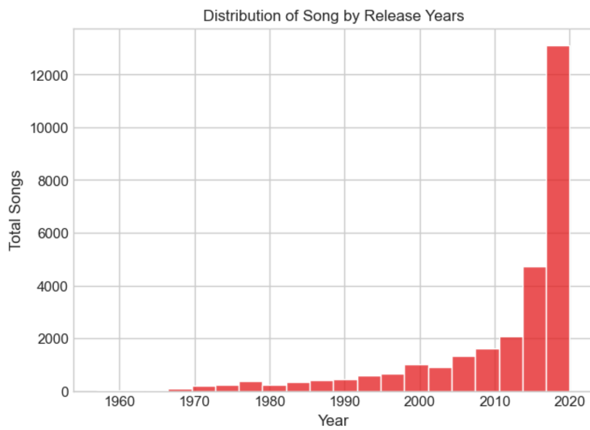


The correlation matrix highlights several interesting relationships between the variables:

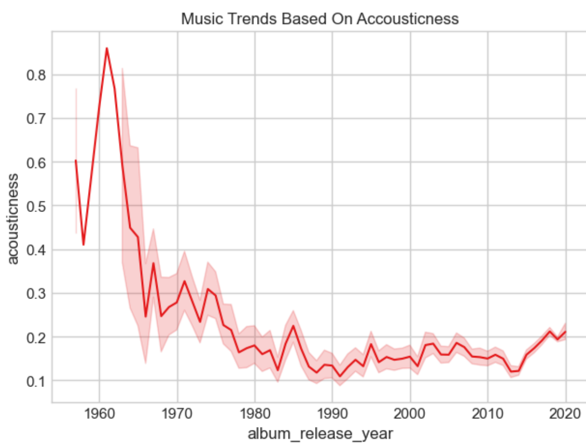
- **Popularity and energy**: A moderate positive correlation, suggesting that more energetic tracks tend to be more popular.
- **Popularity and danceability**: A slightly positive correlation, indicating that more danceable tracks are often more popular.
- **Loudness and energy**: A strong positive correlation, which makes sense as more energetic tracks are generally louder.
- **Instrumentalness and popularity**: A negative correlation, meaning that instrumental tracks are generally less popular.
- **Acousticness and energy**: A negative correlation, which is expected since acoustic tracks tend to be less energetic.

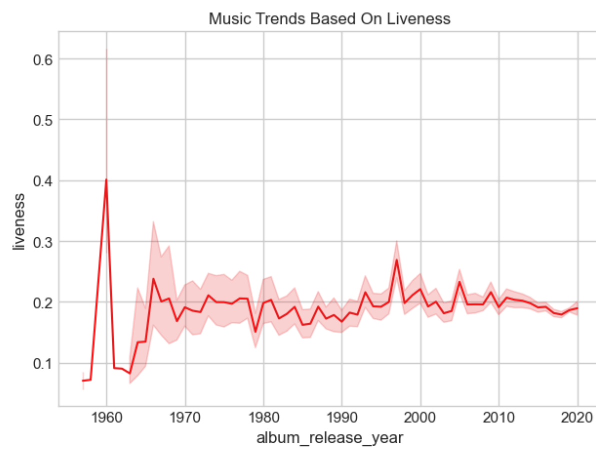
4. Temporal Feature Insights

From the release date column, we extracted three new features: album_release_year, album_release_month, and album_release_day. These allowed us to explore how music characteristics have evolved over time.



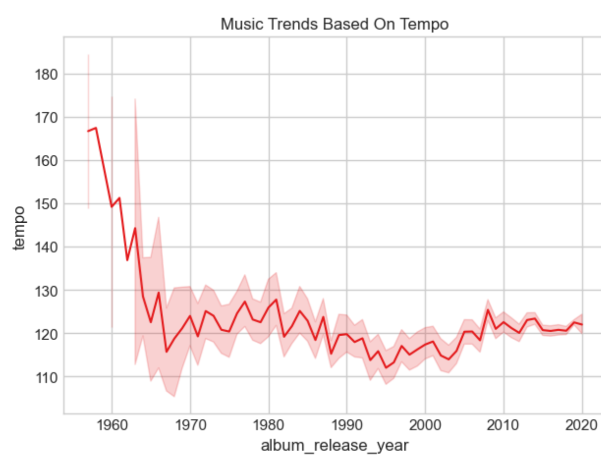
We visualized the number of songs released each year and observed that most of the dataset covers releases between **2000 and 2020**, with a notable increase during the **2010s**, likely due to the rise of digital music platforms.





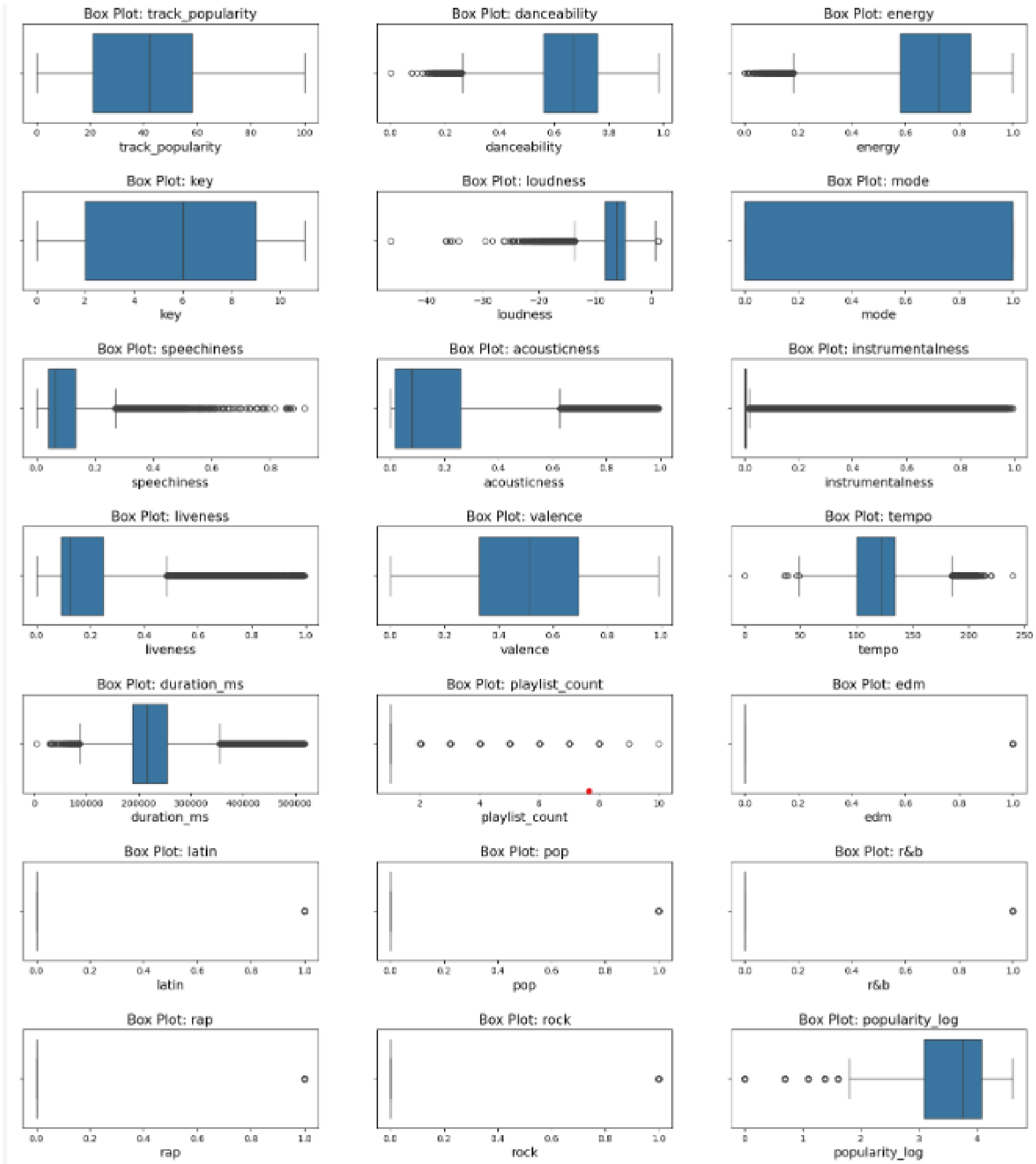
We also plotted the average evolution of key audio features:

- **Acousticness** has decreased steadily, suggesting a shift from acoustic to more electronic music styles.
- **Tempo** has remained relatively stable over time.
- **Liveness** shows minor fluctuations, potentially linked to the increased presence of live recordings in streaming content.



These temporal trends give us useful context and confirm that music production has evolved significantly in the past two decades.

5. Boxplot Summary of All Features



This collection of boxplots gives a global view of the distribution and outliers present across key features.

- **track_popularity** is right-skewed, with many low values and a few extreme hits.

- Features like **tempo**, **speechiness**, **duration_ms**, and **playlist_count** exhibit numerous **outliers**.
- **Acousticness**, **liveness**, and **instrumentalness** show wide distributions, indicating variability across genres.
- Categorical one-hot encoded features (like pop, rap, rock, etc.) are binary and show expected dispersion.

This boxplot matrix helped visualize the **spread, skewness, and extreme values** in the data and informed our decisions for outlier treatment

6. Modeling & Performance

We tested multiple models:

A. Regression:

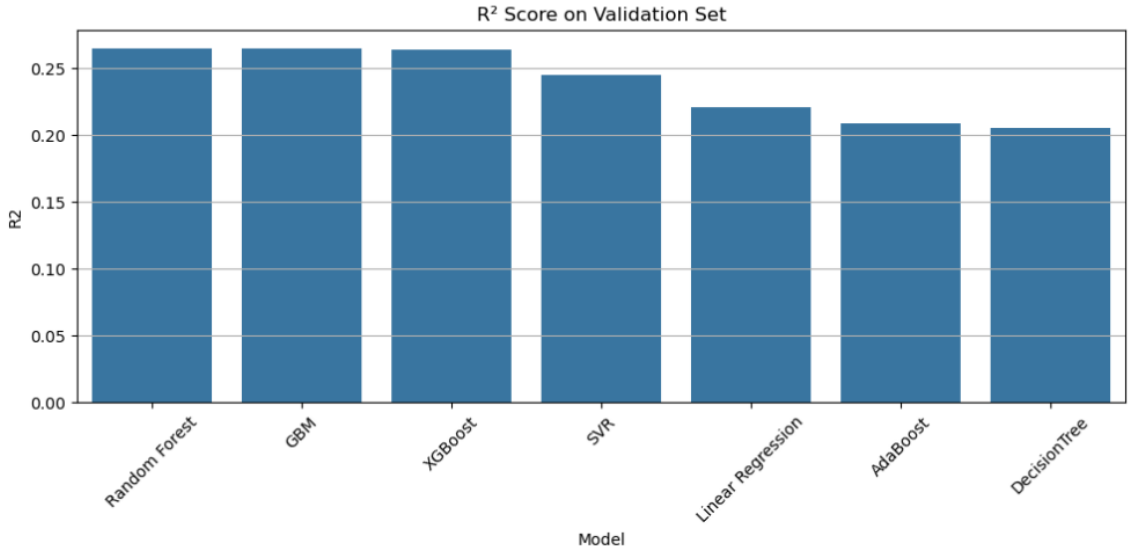
1) Regression Model Comparison

	Model	MSE	RMSE	R2	MAE
2	Random Forest	410.666535	20.264909	0.264877	16.610849
4	GBM	410.833813	20.269036	0.264578	16.528336
6	XGBoost	411.461148	20.284505	0.263455	16.559014
5	SVR	421.783433	20.537367	0.244977	16.149145
0	Linear Regression	435.345271	20.864929	0.220700	17.143643
3	AdaBoost	442.194782	21.028428	0.208439	17.580999
1	DecisionTree	444.059894	21.072729	0.205100	17.253170

This table compares the performance of several regression models in predicting the numerical popularity score of tracks.

- The best R^2 achieved was around **0.264**, using **Random Forest** and **XGBoost**, indicating that less than 27% of the variance in popularity is explained by the model.
- **Random Forest** was selected as the best model based on a balance between RMSE and MAE on the final test set.
- Despite tuning and preprocessing, regression results remained limited, suggesting that **popularity is influenced by external, non-audio factors**.

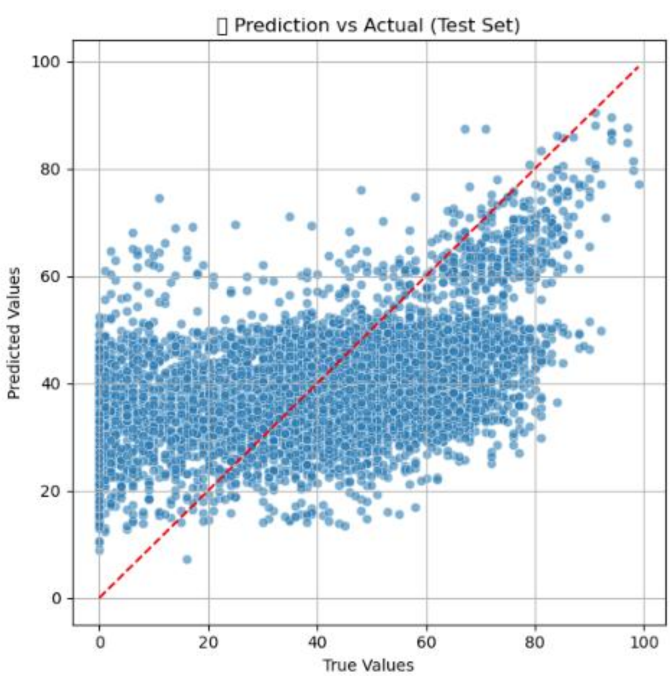
2) R^2 Score by Model (Validation Set)



This bar plot compares the **R^2 scores** of various regression models on the validation set.

- The **Random Forest**, **GBM**, and **XGBoost** models performed the best, each reaching an R^2 slightly above **0.26**.
- **Linear Regression**, **AdaBoost**, and **DecisionTree** models performed worse, confirming that ensemble methods are better suited for this task.

3) Prediction vs Actual (Regression – Test Set)



This scatter plot compares the predicted popularity scores to the actual values on the test set.

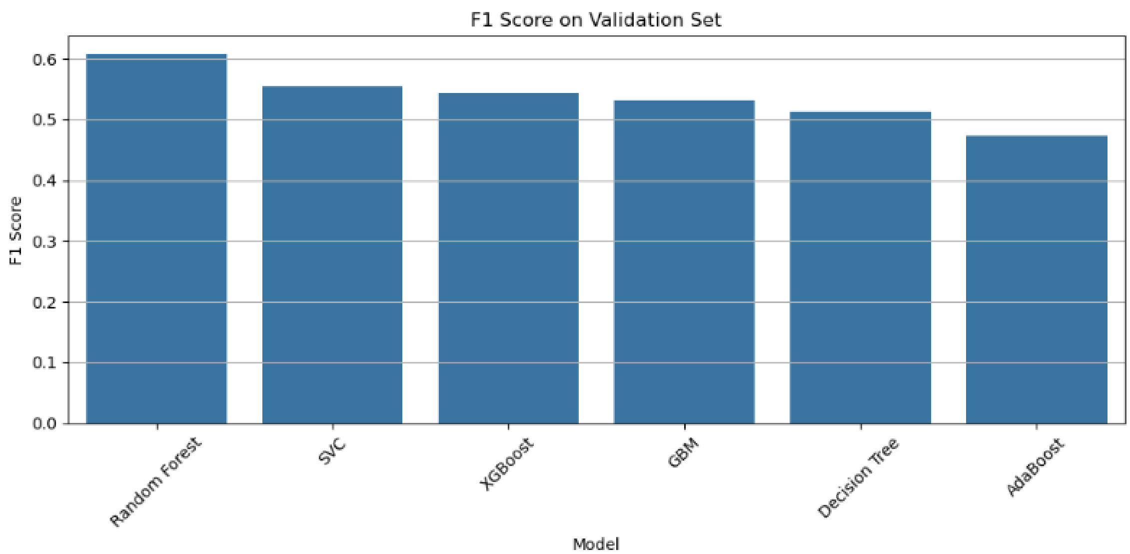
- Each point represents a track. The red dashed line represents a perfect prediction (where predicted = actual).
- While there is a general upward trend, we observe a **wide spread around the diagonal**, especially at lower popularity scores.
- This visual confirms that the model captures general trends but struggles with precise predictions, especially for outlier tracks.

These results led to the decision to reframe the task as a **binary classification problem**.

B. Classification (popularity ≥ 50):

To improve prediction accuracy, we reframed the problem as a binary classification task: determining whether a track is 'popular' (popularity ≥ 50) or not. We evaluated multiple classifiers, including Random Forest, XGBoost, LightGBM, SVC, AdaBoost, Decision Tree, and GBM, using metrics such as F1 Score, Recall, and ROC AUC.

1) Bar plot : F1 Score on Validation Set



This bar chart displays the **F1 score** of each classification model on the validation set.

- **Random Forest** achieved the best F1 score, confirming its strength in balancing **precision and recall**.
- **SVC** and **XGBoost** also performed well, while **AdaBoost** had the weakest performance.
- F1 score is a key metric in this context because it balances the trade-off between:
  - **Precision**: not falsely predicting popularity
  - **Recall**: not missing truly popular songs

F1 score guided the model selection process. Since Random Forest showed the best balance between capturing true positives and avoiding false alarms, it was chosen as one of the final models.

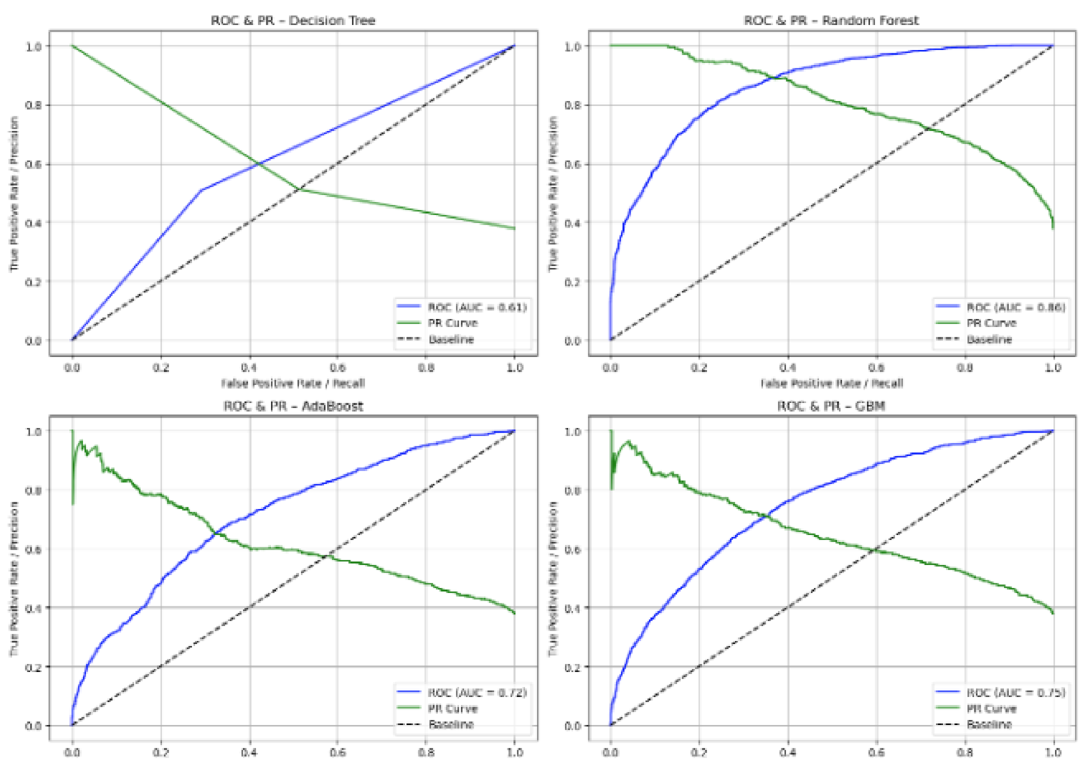
2) Model Comparison: Key Metrics

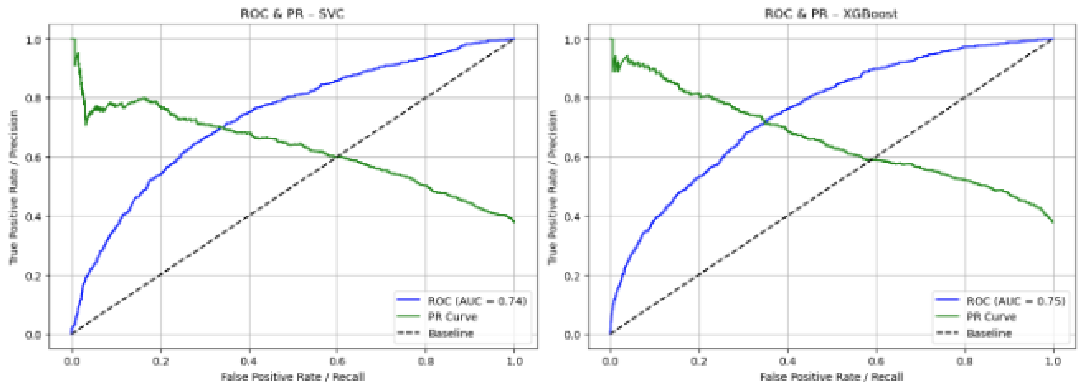
	Model	Accuracy	Precision	Recall	F1	ROC AUC
1	Random Forest	0.694	0.591	0.625	0.607	0.748
4	SVC	0.704	0.646	0.486	0.554	0.735
5	XGBoost	0.706	0.661	0.460	0.543	0.754
3	GBM	0.699	0.648	0.451	0.532	0.747
0	Decision Tree	0.632	0.515	0.510	0.512	0.608
2	AdaBoost	0.674	0.610	0.386	0.473	0.715

This table summarizes the performance of several classifiers using key metrics: **Accuracy**, **Precision**, **Recall**, **F1 score**, and **ROC AUC**.

- **Random Forest** and **XGBoost** yielded the best balance of metrics.
  - **XGBoost** achieved the highest **Precision (0.861)** and **ROC AUC (0.754)**.
  - **Random Forest** had a higher **Recall (0.625)** and a strong F1 score (0.607).
- **SVC** and **GBM** performed moderately well.
- **Decision Tree** and **AdaBoost** had lower F1 scores and AUCs, making them less suitable.

3) Receiver Operating Characteristic (ROC) & Precision-Recall (PR)





This plot shows both **ROC curves (blue)** and **Precision-Recall curves (green)** for all classifiers.

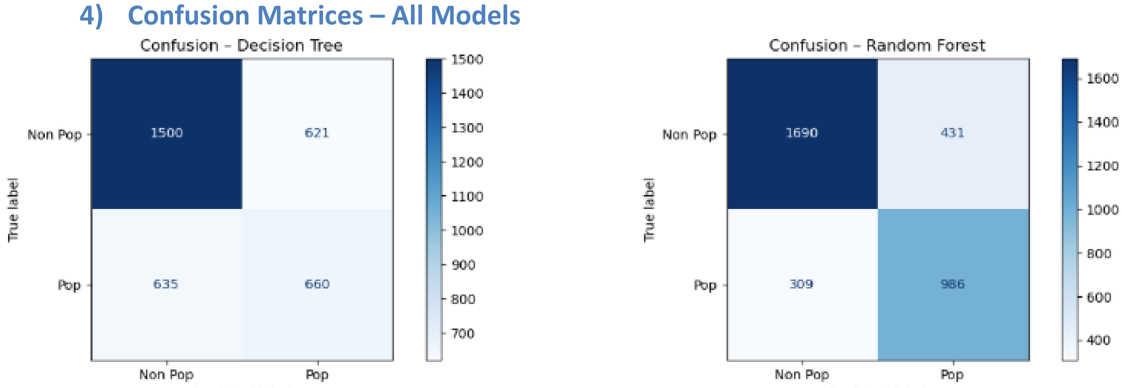
Interpretation:

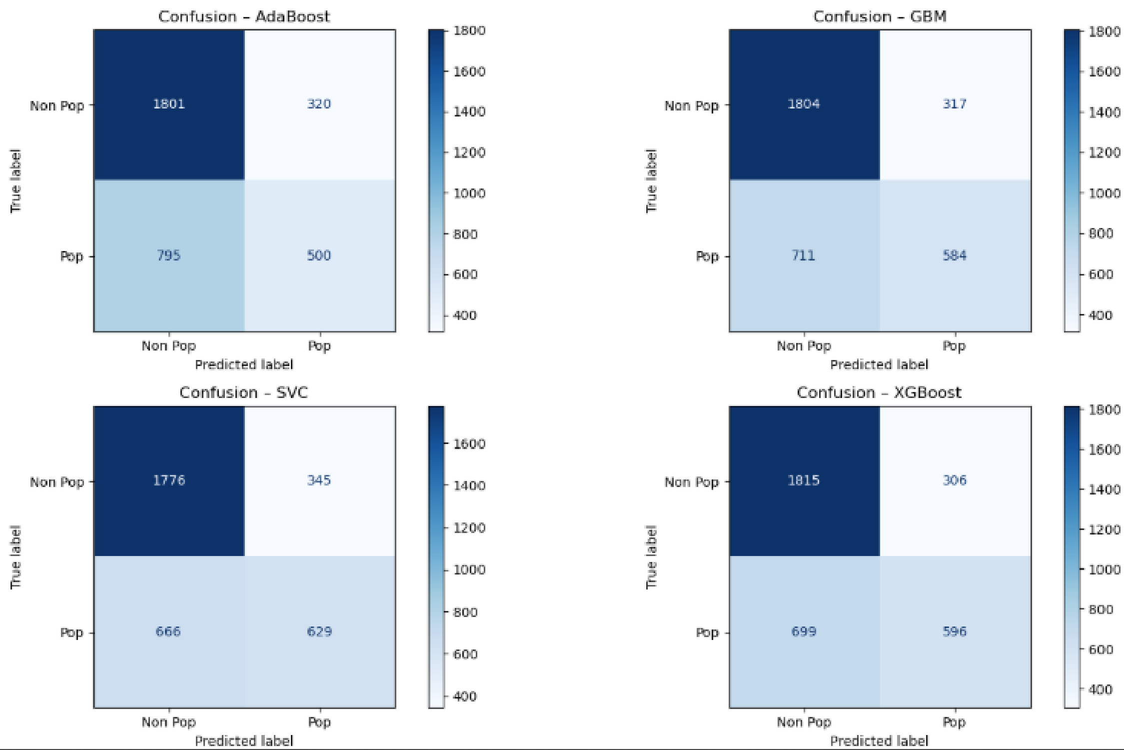
- The **ROC curve** evaluates how well the model distinguishes between classes regardless of the decision threshold.
    - AUC closer to 1 = better performance.
- The **Precision-Recall curve** highlights how well the model performs under class imbalance.

Insights:

- **Random Forest**, **XGBoost**, and **SVC** have the **highest ROC AUC values (~0.75)** and strong PR curves — confirming their robustness.
- **Decision Tree** shows the weakest performance (low ROC and poor PR shape).
- **PR curves** are especially useful in this case, since our positive class (popular songs) is less represented than non-popular ones.

4) Confusion Matrices – All Models





This grid of confusion matrices allows a **visual comparison of classification performance** across all models. Each matrix shows how many songs were:

- **Correctly classified** (top-left and bottom-right = True Negatives & True Positives)
- **Incorrectly classified** (top-right = False Positives, bottom-left = False Negatives)

Random Forest and **XGBoost** have the **highest number of true positives** (bottom-right), indicating strong ability to detect popular songs.

- **AdaBoost** and **Decision Tree** misclassify more popular songs (more false negatives).
- **SVC** achieves a good balance between FP and FN.

These matrices complement the **F1 score and ROC AUC metrics** and help us understand **not just how well a model performs, but where it makes mistakes**.

5) Final Test Set Results– Random Forest:

The Random Forest model, selected for its balanced performance during validation, achieved the following metrics on the final test set:

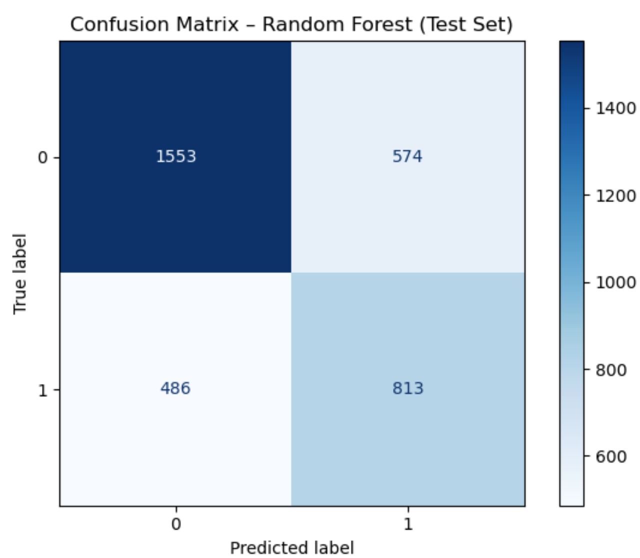
```
✅ Final test performance for Random Forest:
{'Accuracy': 0.691, 'Precision': 0.586, 'Recall': 0.626, 'F1': 0.605, 'ROC AUC': 0.752}
```

- **Accuracy (0.691)**: 69.1% of total predictions were correct. However, this metric can be misleading if classes are imbalanced.
- **Precision (0.580)**: Among all the tracks predicted as “popular”, 58% were actually popular. Useful when we want to avoid false positives.
- **Recall (0.626)**: The model correctly identified 62.6% of all truly popular tracks. Important if we want to avoid missing hits.
- **F1 Score (0.605)**: Harmonic mean of precision and recall – it balances both and gives a more comprehensive view.
- **ROC AUC (0.752)**: Measures the model’s ability to distinguish between popular and non-popular tracks. AUC of 0.75 indicates good separation.

These results show that the model offers a balanced trade-off between precision and recall and performs well at distinguishing between the two classes.

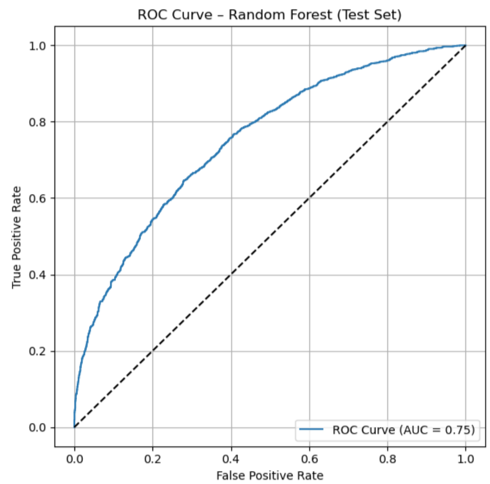
6) Confusion Matrix – Random Forest

This matrix highlights how the Random Forest model performed in terms of true/false positive and negative classifications on the test set :



- The model correctly identified **813 popular songs** and **1553 non-popular ones**.
- It **misclassified 486 popular tracks** as not popular (false negatives).
- It also **wrongly flagged 574 songs as popular** that were not (false positives).
- This balance of errors explains the **moderate precision and recall**, and supports the F1 score of ~0.60.

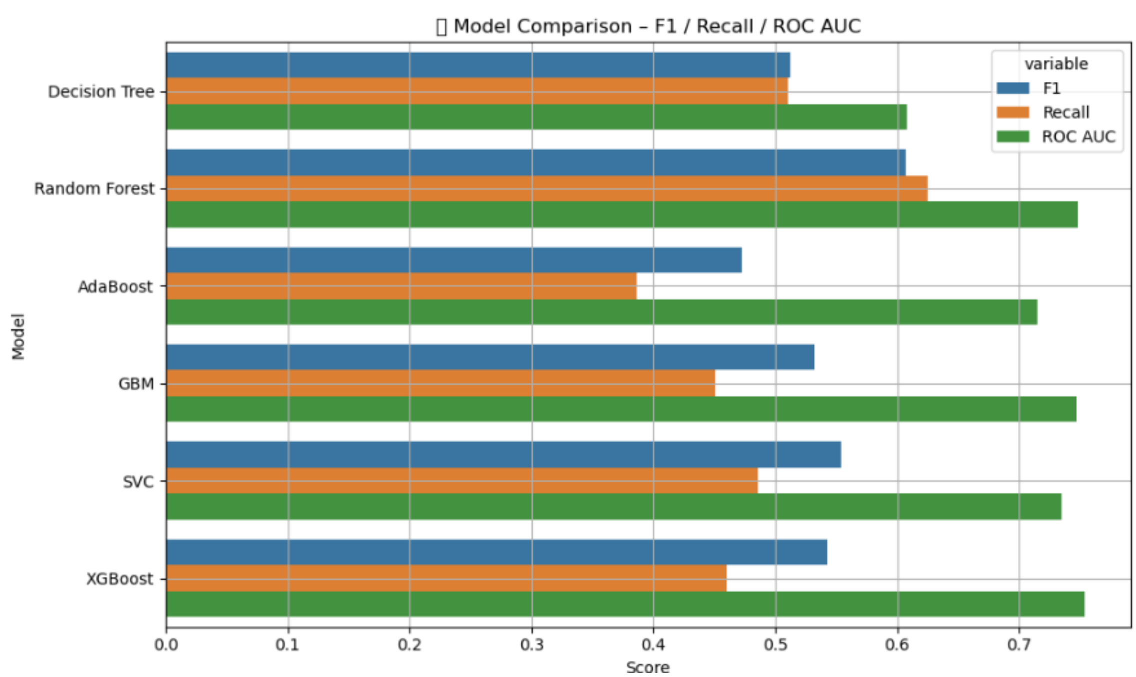
7) ROC Curve – Random Forest



The ROC curve for the final Random Forest model shows an AUC of 0.75, indicating strong class discrimination performance.

8) Model Performance Summary – F1, Recall, ROC AUC

	Model	Accuracy	Precision	Recall	F1	ROC AUC
1	Random Forest	0.694	0.591	0.625	0.607	0.748
4	SVC	0.704	0.646	0.486	0.554	0.735
5	XGBoost	0.706	0.661	0.460	0.543	0.754
3	GBM	0.699	0.648	0.451	0.532	0.747
0	Decision Tree	0.632	0.515	0.510	0.512	0.608
2	AdaBoost	0.674	0.610	0.386	0.473	0.715



This chart combines F1 Score, Recall, and ROC AUC for all models, providing a quick comparison of their global performance. It confirms that Random Forest, XGBoost, and SVC offer the most balanced and effective classification results. This visualization supports the final model selection process and offers an at-a-glance summary for stakeholders