What makes a song “Popular”?

Background & motivation

Why song Popularity?

Want to understand what attributes “popular” songs had

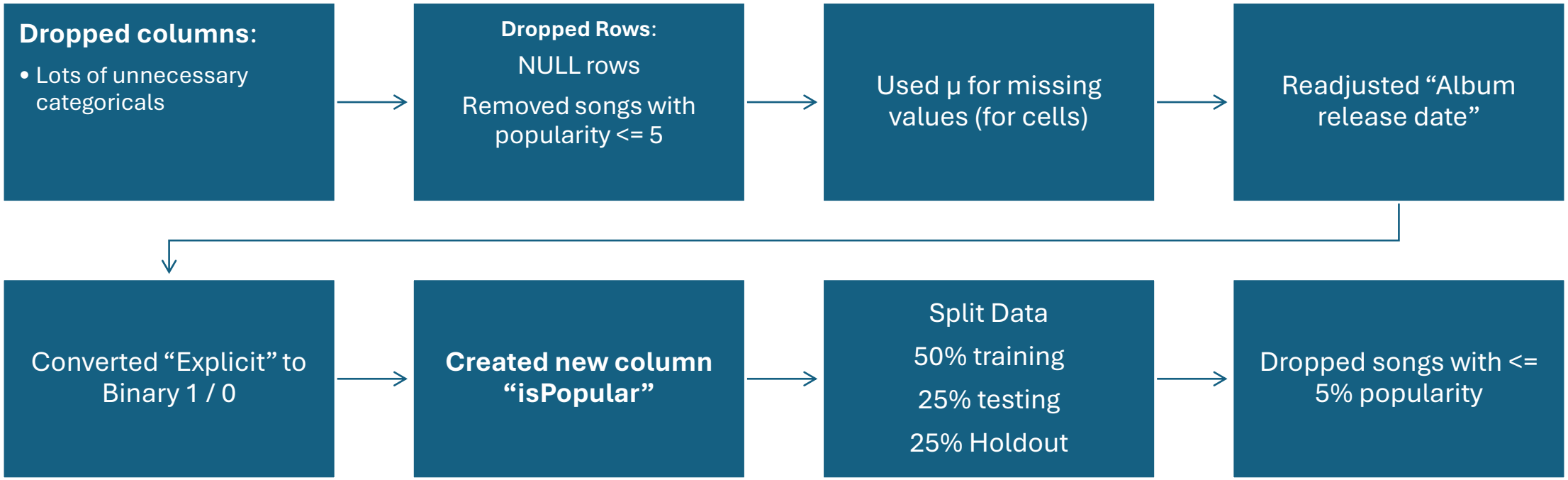
Goal?

Find out which features/observations are important for a song’s success using **Logistic Regression** & **Random Forest Classification**

Data Source

- Dataset used:
  - Kaggle: Top 10,000 Spotify Songs (1960–now)
    - Source: https://www.kaggle.com/datasets/joebeachcapital/top-10000-spotify-songs-1960-now

Data Preprocessing



1st Technique: Logistic Regression

Why?

Classify (Log) rather than **predict** (Linear)

How?

Added constants in xTraining for **Logit**

- **Logit** is good for P-Values

Ran thresholds for “isPopular”

- 15-60% popularity intervals

Results (51%)

- Used backwards selection
  - Very balanced model

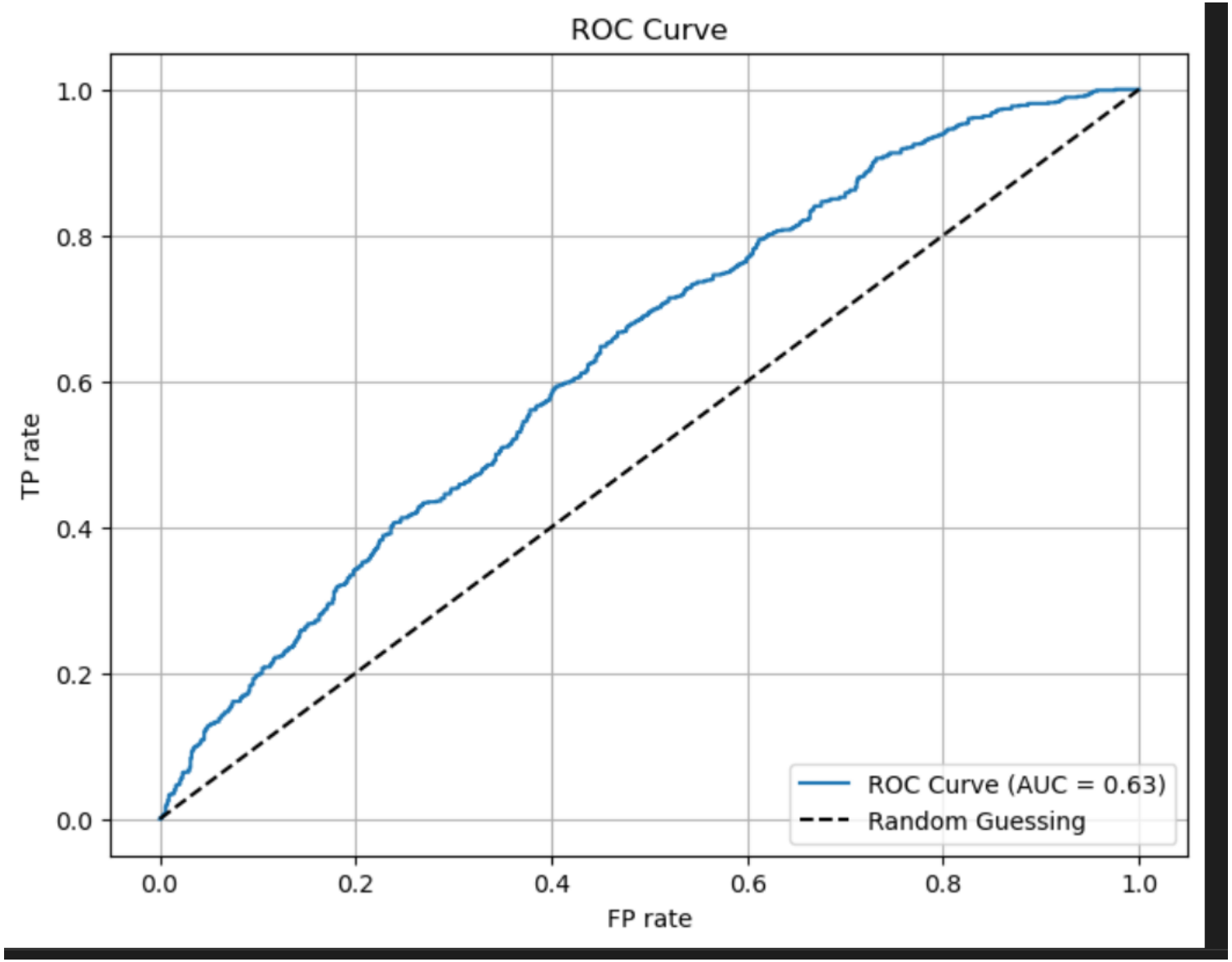
```
[[557 303]
 [405 421]]
Accuracy: 0.5800711743772242
              precision    recall  f1-score   support

           0       0.58      0.65      0.61       860
           1       0.58      0.51      0.54       826

    accuracy                           0.58      1686
   macro avg       0.58      0.58      0.58      1686
weighted avg       0.58      0.58      0.58      1686
```

```
Optimization terminated successfully.
         Current function value: 0.666215
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:              isPopular   No. Observations:                 3372
Model:                          Logit   Df Residuals:                     3361
Method:                           MLE   Df Model:                           10
Date:                Sun, 27 Apr 2025   Pseudo R-squ.:                0.03809
Time:                        15:09:37   Log-Likelihood:                -2246.5
converged:                       True   LL-Null:                       -2335.4
Covariance Type:            nonrobust   LLR p-value:                6.364e-33
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             45.4963      5.670      8.024      0.000      34.384      56.609
Album Release Date  -0.0225      0.003     -8.071      0.000      -0.028      -0.017
Track Duration (ms) 3.112e-06   7.55e-07      4.124      0.000    1.63e-06    4.59e-06
Explicit            0.7112      0.165      4.306      0.000       0.387       1.035
Danceability        1.5036      0.303      4.958      0.000       0.909       2.098
Energy             -1.0003      0.328     -3.045      0.002      -1.644      -0.357
Loudness            0.1104      0.018      6.276      0.000       0.076       0.145
Acousticness       -0.5439      0.196     -2.775      0.006      -0.928      -0.160
Instrumentalness   -0.8332      0.334     -2.496      0.013      -1.488      -0.179
Liveness           -0.8462      0.249     -3.398      0.001      -1.334      -0.358
Valence            -0.7288      0.197     -3.693      0.000      -1.116      -0.342
==============================================================================
```

ROC curve



Note: Curve remained consistent, no matter the accuracy & threshold

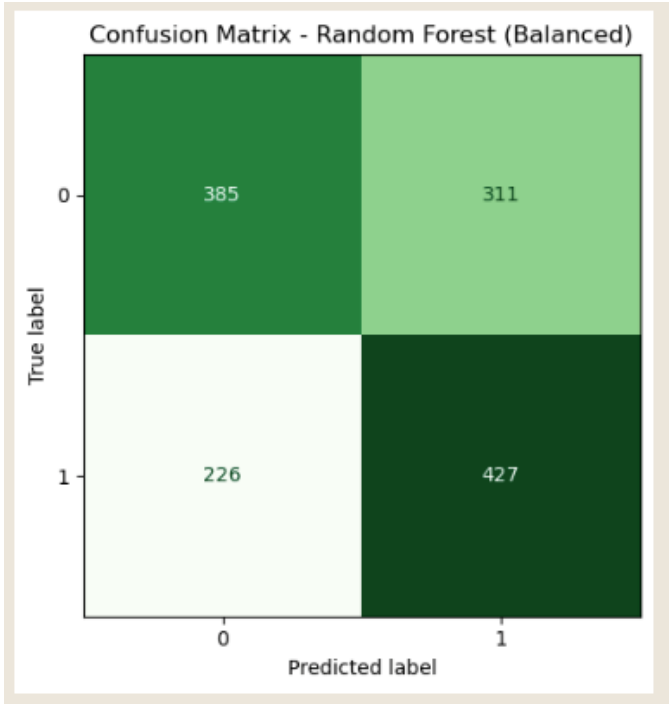
2nd Technique: Random Forest Classification

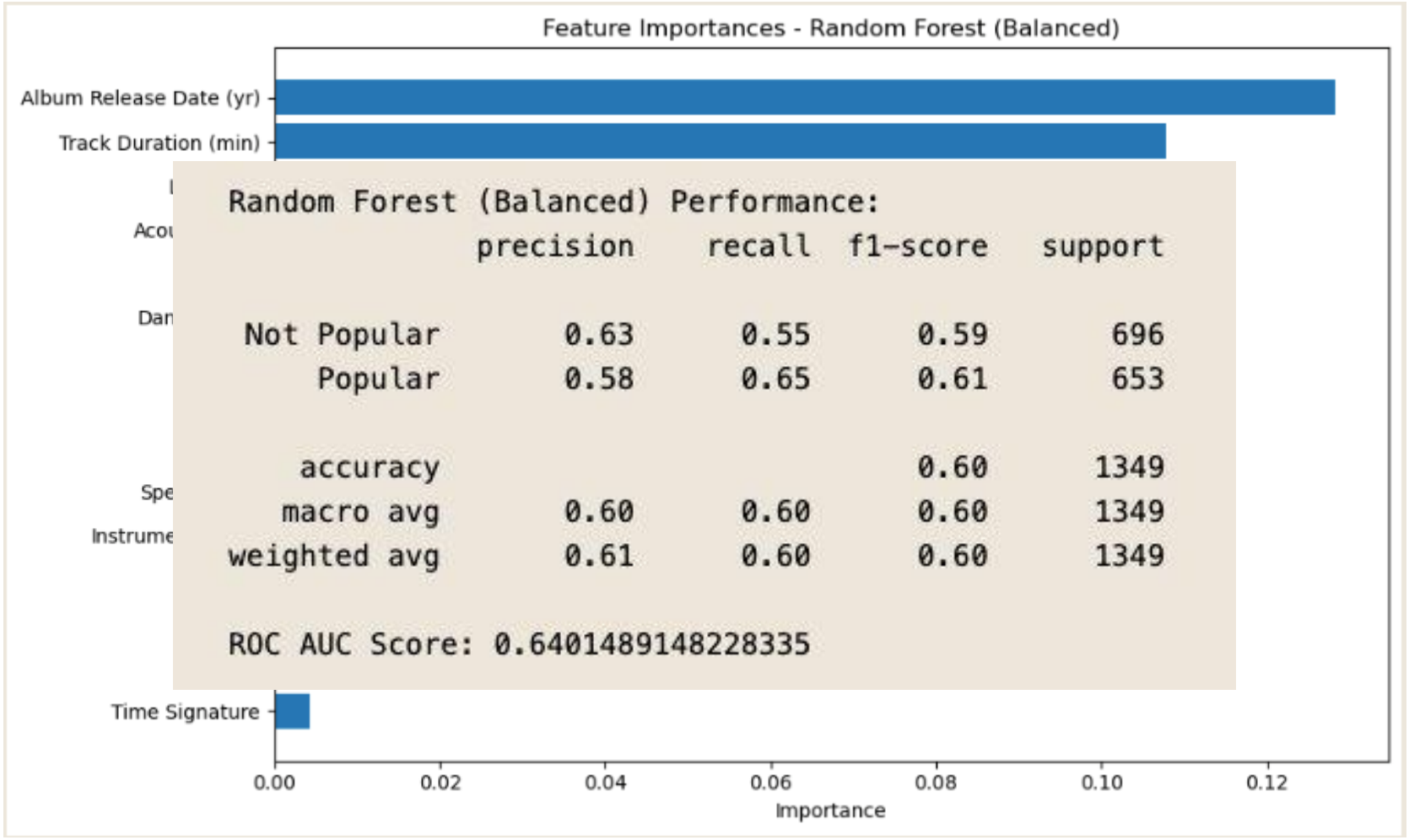
Why?
• Non Linear Relationships • Best possible ratio performance • Ensembles learning

How?
• Tune Hyperparameter • Showed feature importances based on percentage • Reduced variance

```
rf_audio_balanced = RandomForestClassifier(
    n_estimators=200,
    max_depth=10,
    random_state=42,
    class_weight='balanced'
)
rf_audio_balanced.fit(X_train, y_train)
```

Results (51%)



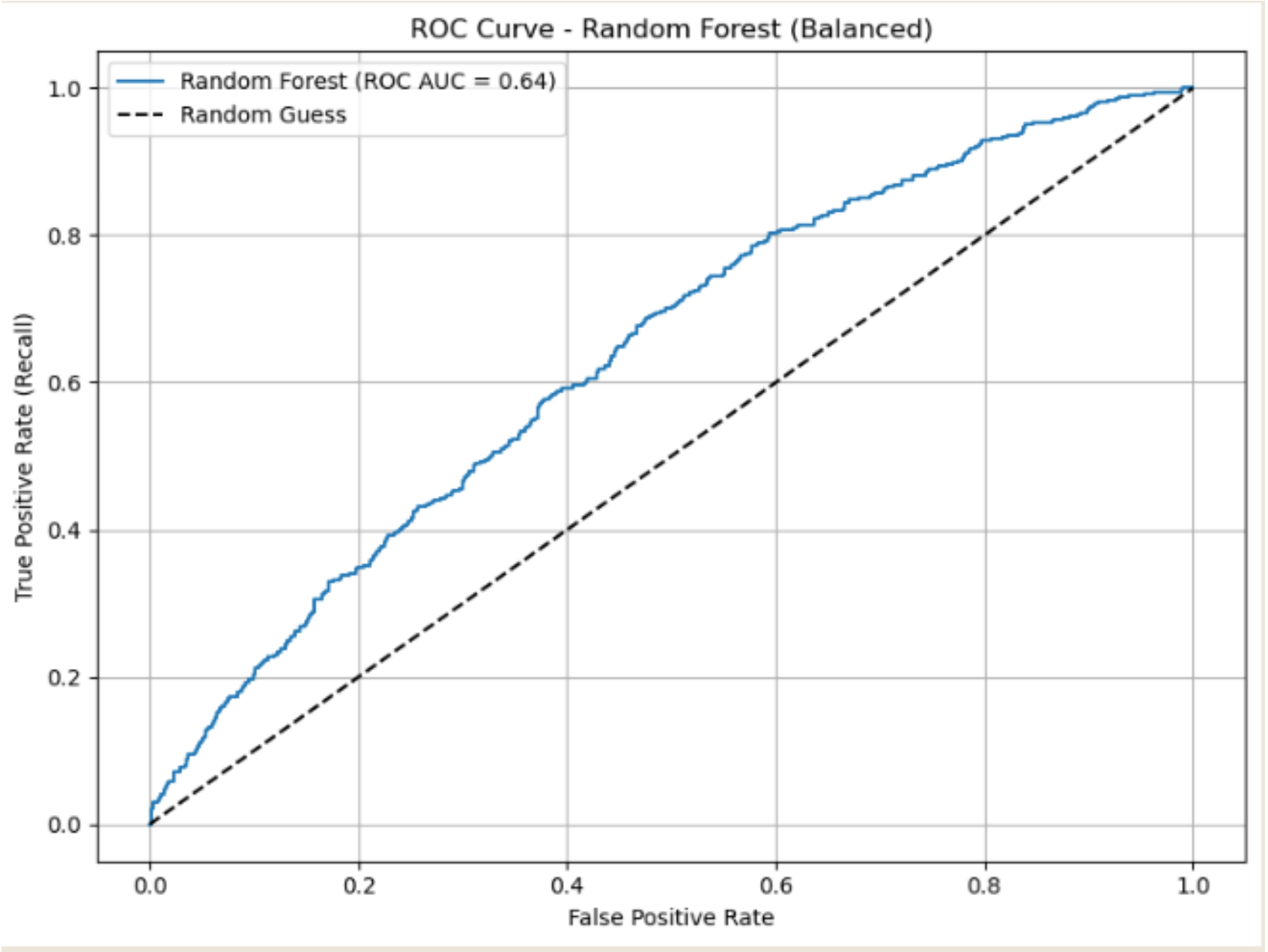


Random Forest (Balanced) Performance:

	precision	recall	f1-score	support
Not Popular	0.63	0.55	0.59	696
Popular	0.58	0.65	0.61	653
accuracy			0.60	1349
macro avg	0.60	0.60	0.60	1349
weighted avg	0.61	0.60	0.60	1349

ROC AUC Score: 0.6401489148228335

ROC curve


ROC Curve - Random Forest (Balanced)
Random Forest (ROC AUC = 0.64)
Random Guess
True Positive Rate (Recall)
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Key features for Each Model

Ranking	Logistic Regression: p-value	Random Forest: Importance Percentage
1	**Album Release Date (year) : 0.000**	**Album Release Date (yr): 0.128307**
2	**Track Duration (ms): 0.000**	**Track Duration (min): 0.107764**
3	Explicit: 0.000	**Loudness: 0.092928**
4	**Danceability: 0.000**	Acousticness: 0.083351
5	**Valence: 0.000**	**Liveness: 0.080210**
6	**Loudness: 0.000**	**Danceability: 0.079675**
7	**Liveness: 0.001**	**Valence: 0.079623**
8	Energy: 0.002	Tempo: 0.074567

Best ranges of Key Features

- Using μ of the chosen column & the popularity column
  - Binned into 5 widths

Agreed Features	Best range
Album Release Date	1968-1983
Track Duration (min)	1.5 – 6.00
Danceability	0.593- 0.79
Loudness	-5.37- -0.358
Valence	0.0- 0.2
Liveness	0.011- 0.205

Conclusion

- Very few technical features affect a song's popularity

- Limitations
  - Very imbalanced dataset (towards unpopular songs)
  - Useless metadata