SPOTIFY

USER CHURN ANALYSIS

Predicting Spotify churn using user engagement behavior to improve retention through early intervention.

By: Soham Sheemar

SPOTIFY, LIKE OTHER SUBSCRIPTION-BASED PLATFORMS, FACES A CRITICAL BUSINESS CHALLENGE: **USER CHURN**. THIS PROJECT AIMS TO BUILD A PREDICTIVE MODEL TO IDENTIFY USERS WHO ARE LIKELY TO STOP USING THE PLATFORM, ENABLING SPOTIFY TO TAKE PREVENTIVE ACTIONS TO IMPROVE RETENTION.

Music
Company

Presentation Template 2022

PROBLEM STATEMENT

Can we accurately predict which Spotify users are at risk of churning using behavioral and engagement data? The goal is to maximize recall for churn prediction so that more at-risk users can be flagged.

DATASET

Source: KaggleA structured dataset of 1,000 Spotify users tracking subscription behavior, engagement metrics, and churn status for retention analysis.

ABOUT

- **User Profile**: Subscription type (Free/Premium), country
- **Behavioral Data**: Daily listening time, playlists, top genre, skips
- **Engagement Metrics**: Support tickets, days since last login
- **Target Variable**: Binary churn indicator (0=active, 1=churned)

DATA CONTAINS

- 1,000 records (rows)
- 10 features (columns)

KEY FEATURES

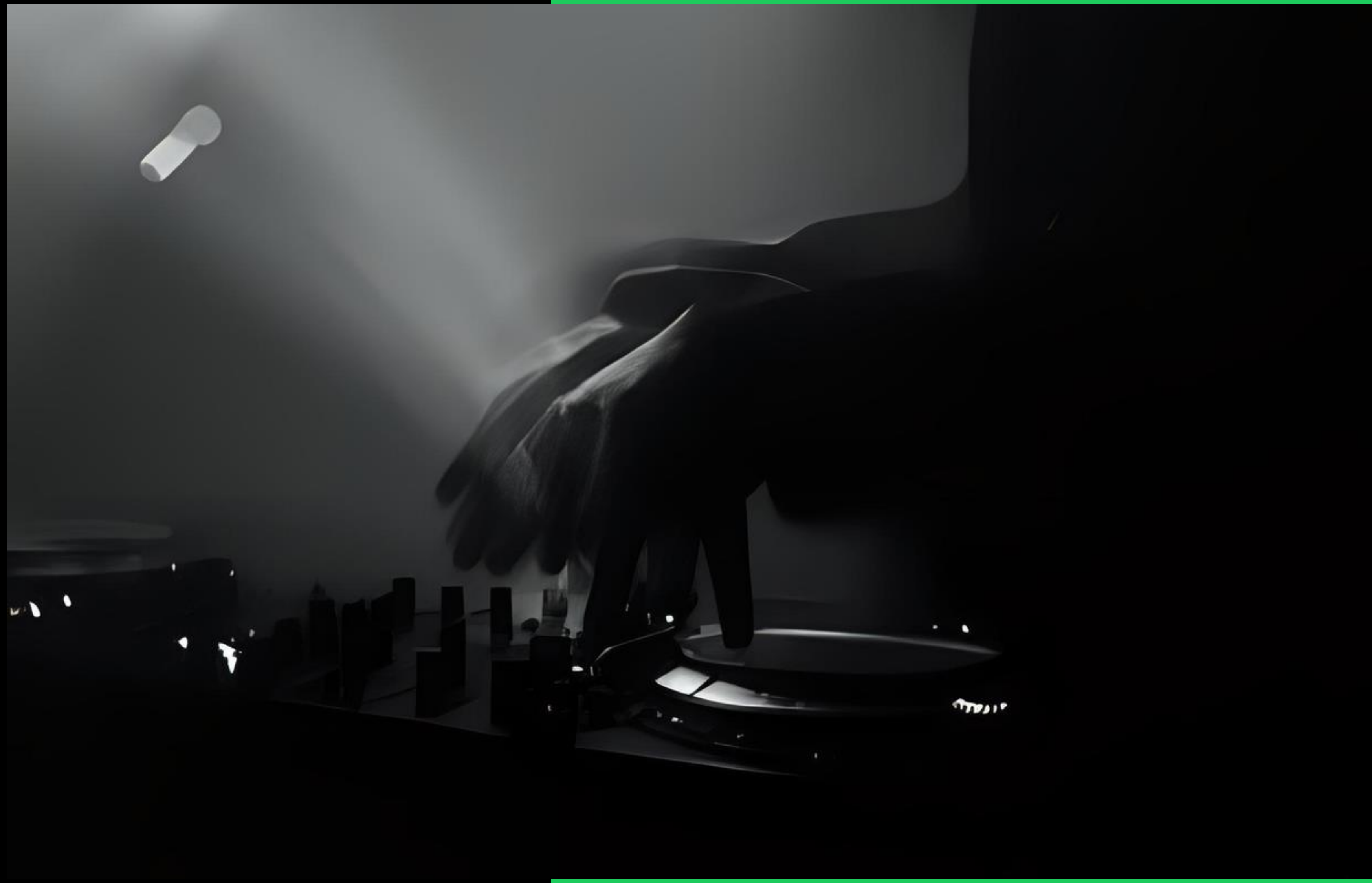
- **avg_daily_minutes**
- **days_since_last_login**
- **num_playlists**
- **engagement_score (engineered)**
- **skips_per_day, support_tickets, session_count, subscription_type**

Total Size
1000

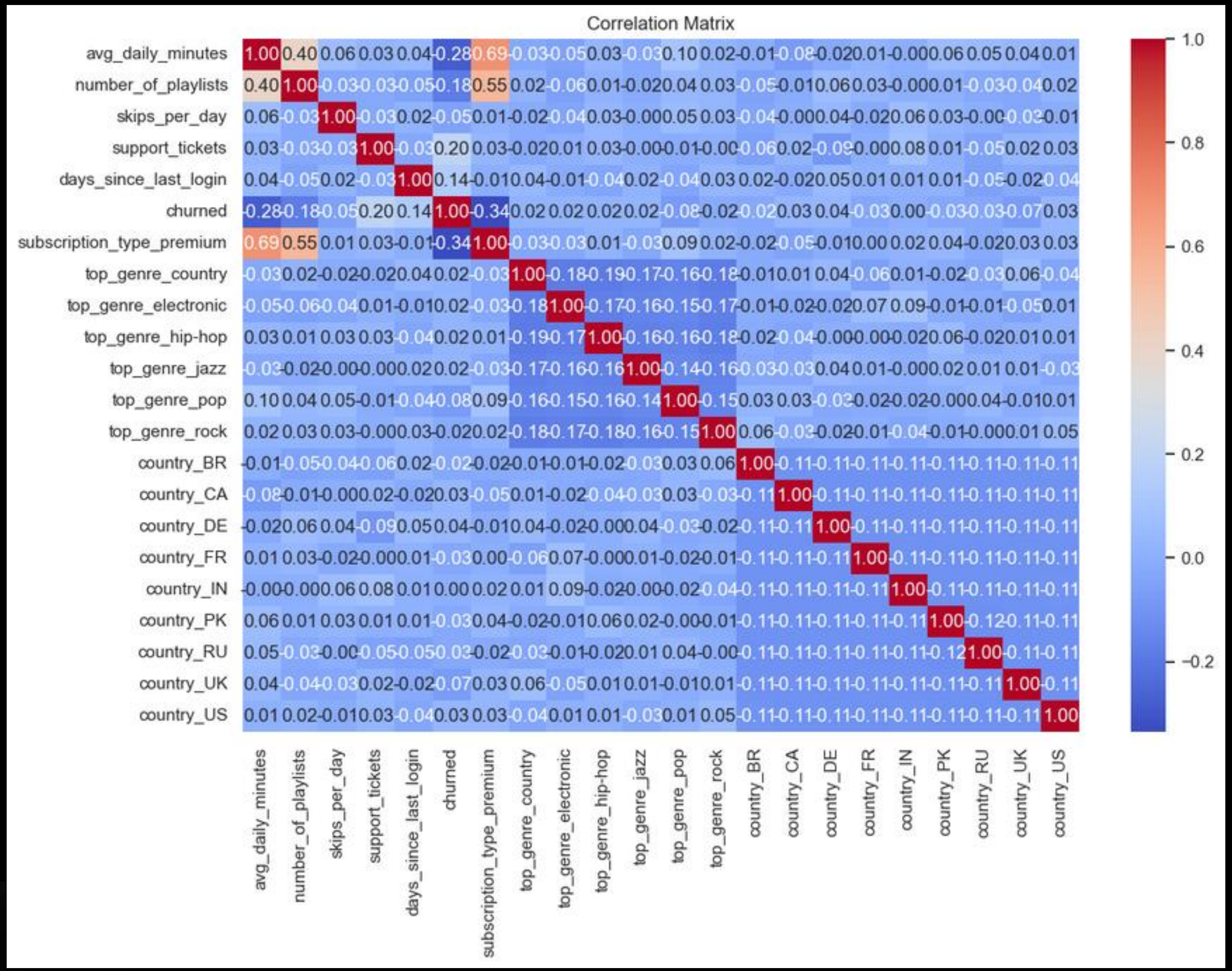
After Cleaning
993

APPROACH

- Data Cleaning: Handled nulls, removed irrelevant IDs, created meaningful features.
- Feature Engineering: Calculated an "engagement_score" combining time spent, sessions, skips, and recency.
- EDA: Visualized correlations, top drivers of churn.
- Modeling:
  - Logistic Regression (class_weight='balanced')
  - Random Forest (with SMOTE and custom threshold)
  - Performance measured via accuracy, precision, recall, f1-score, and ROC AUC.



ANALYSIS

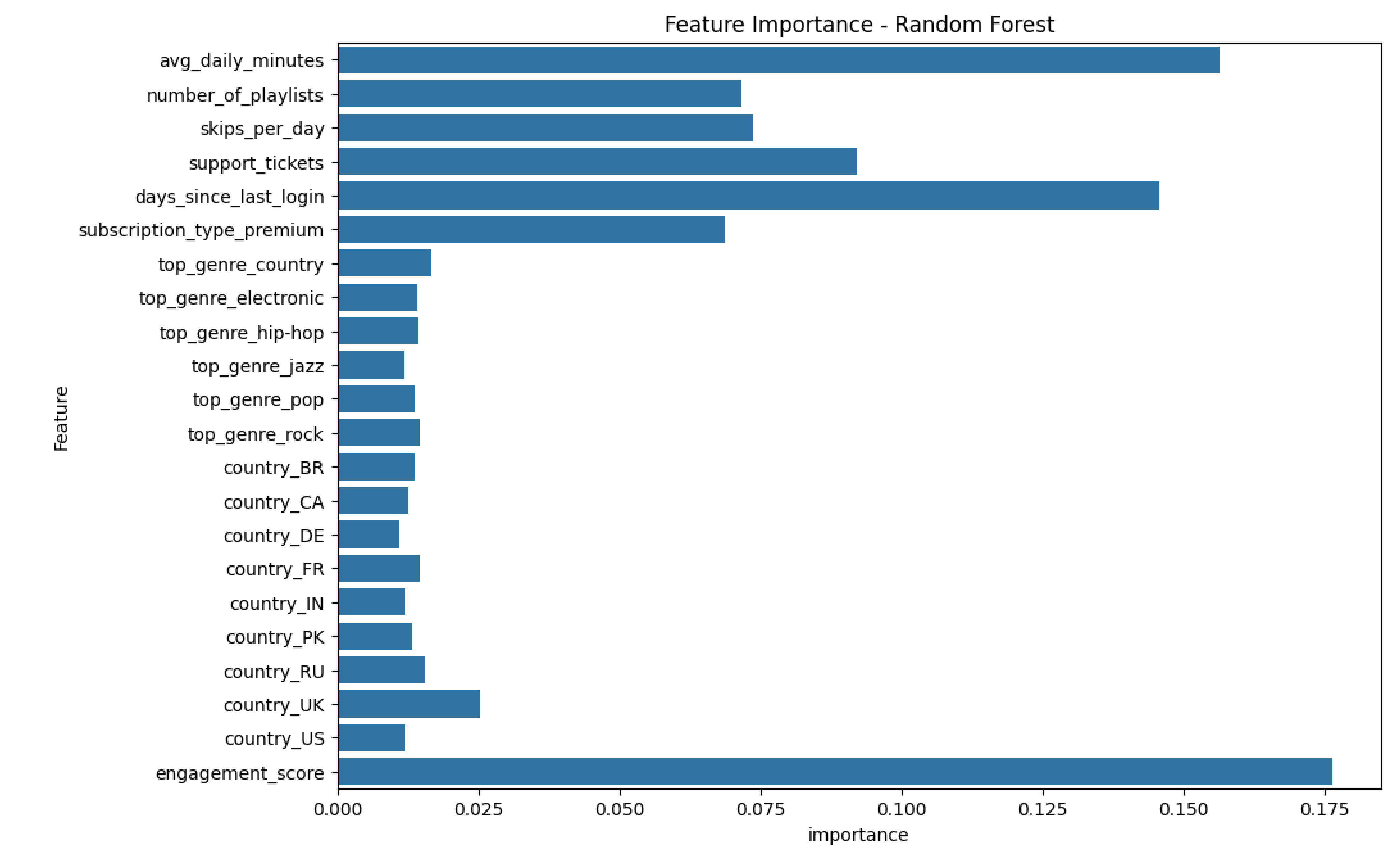


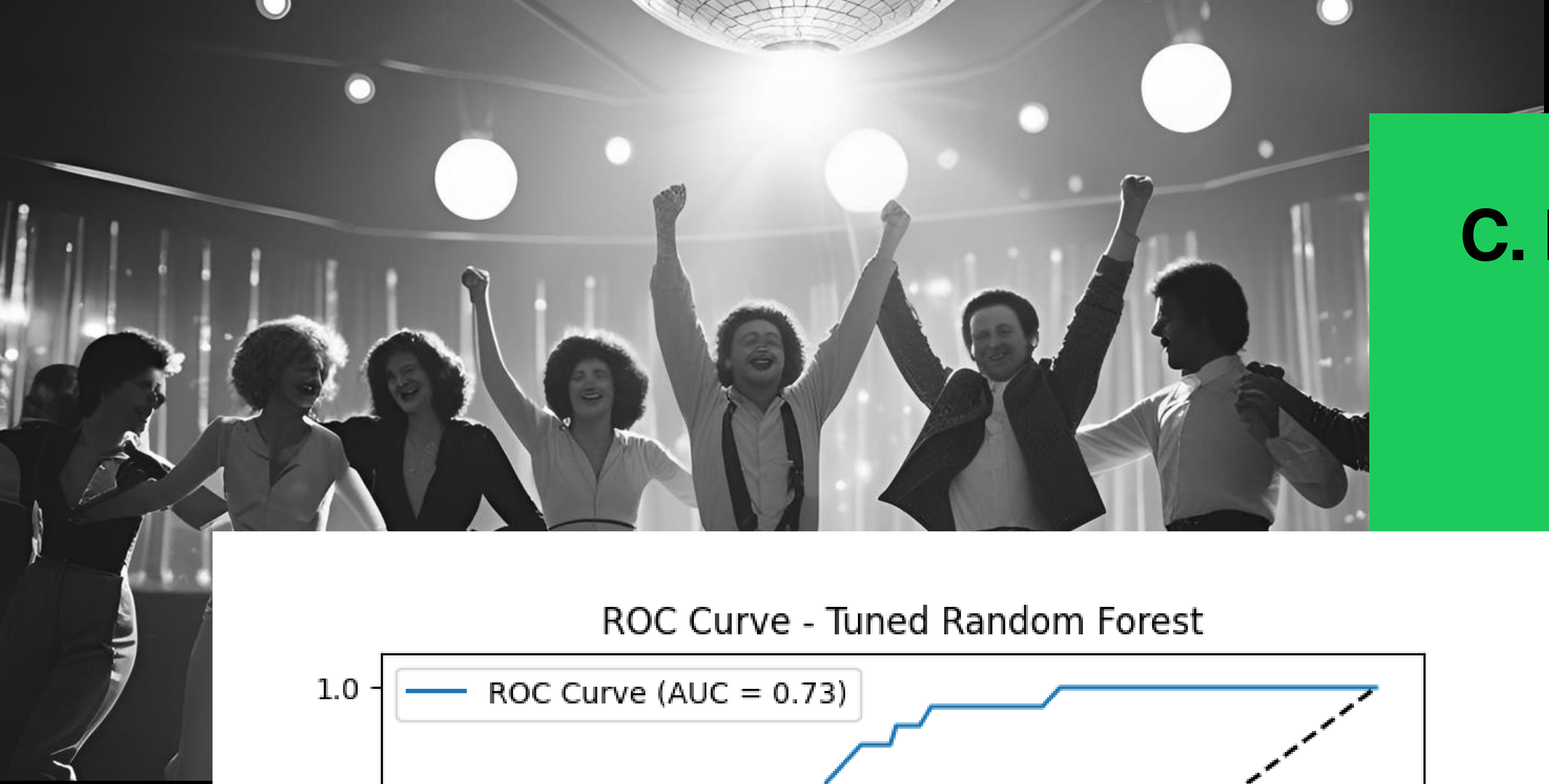
A. CORRELATION MATRIX

- days_since_last_login shows strong positive correlation with churn.
- avg_daily_minutes and engagement_score show strong negative correlation with churn.

B. FEATURE IMPORTANCE (RANDOM FOREST)

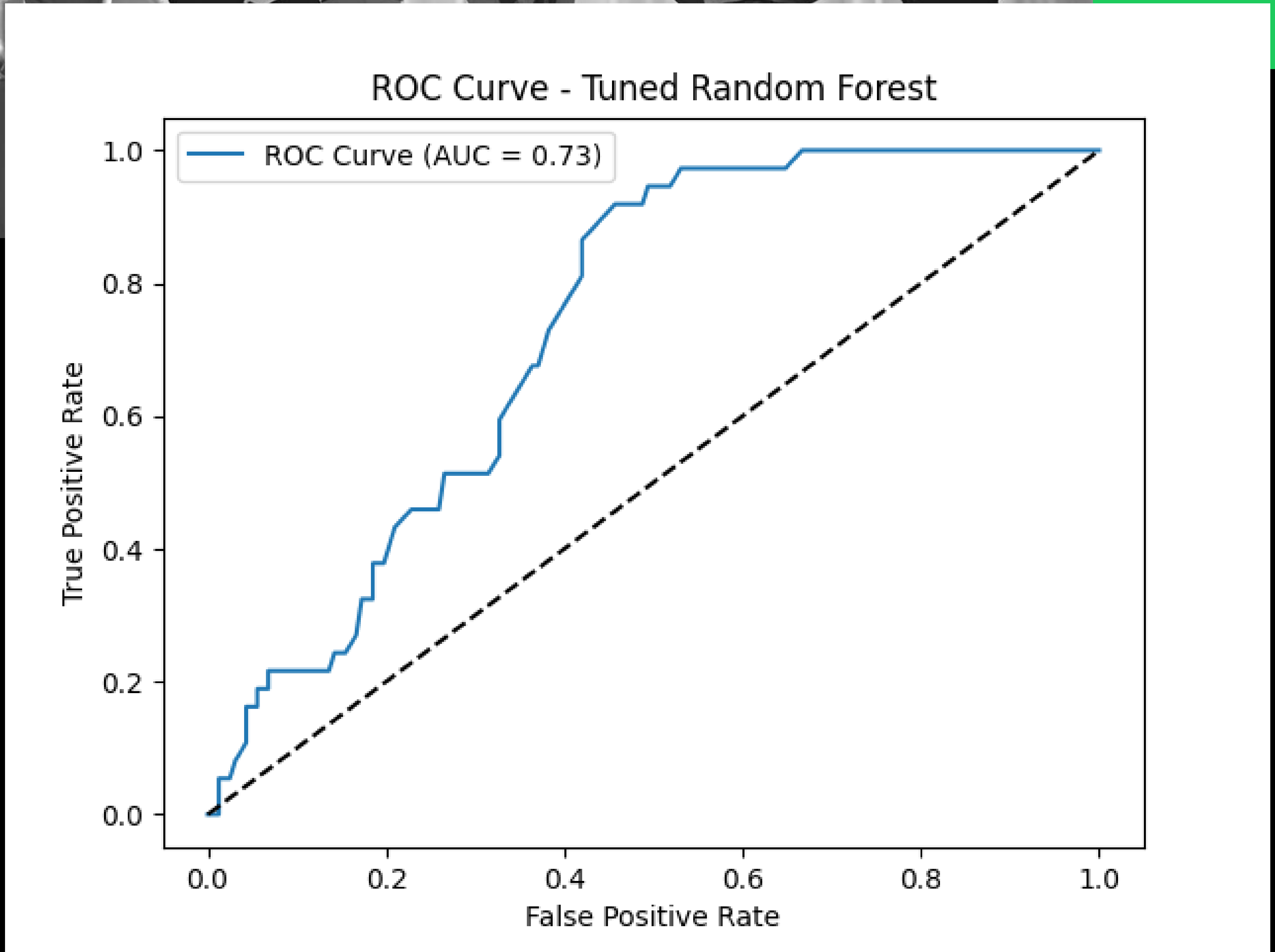
- Top features:
- days_since_last_login
- avg_daily_minutes
- skips_per_day
- engagement_score

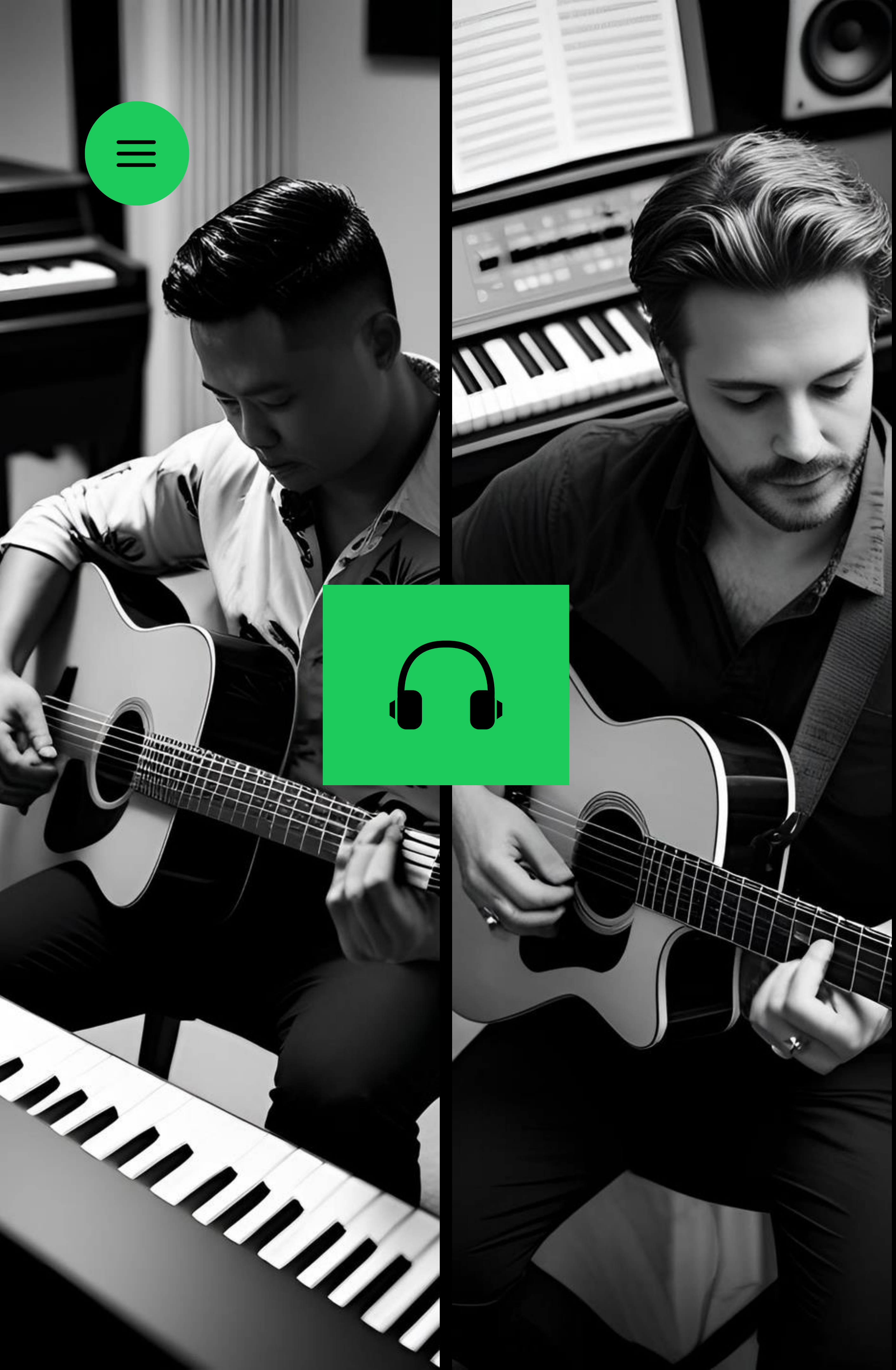




C. ROC CURVE (RANDOM FOREST AUC ~0.79)

- Demonstrates good discrimination between churners and non-churners.





D. MODEL EVALUATION SUMMARY

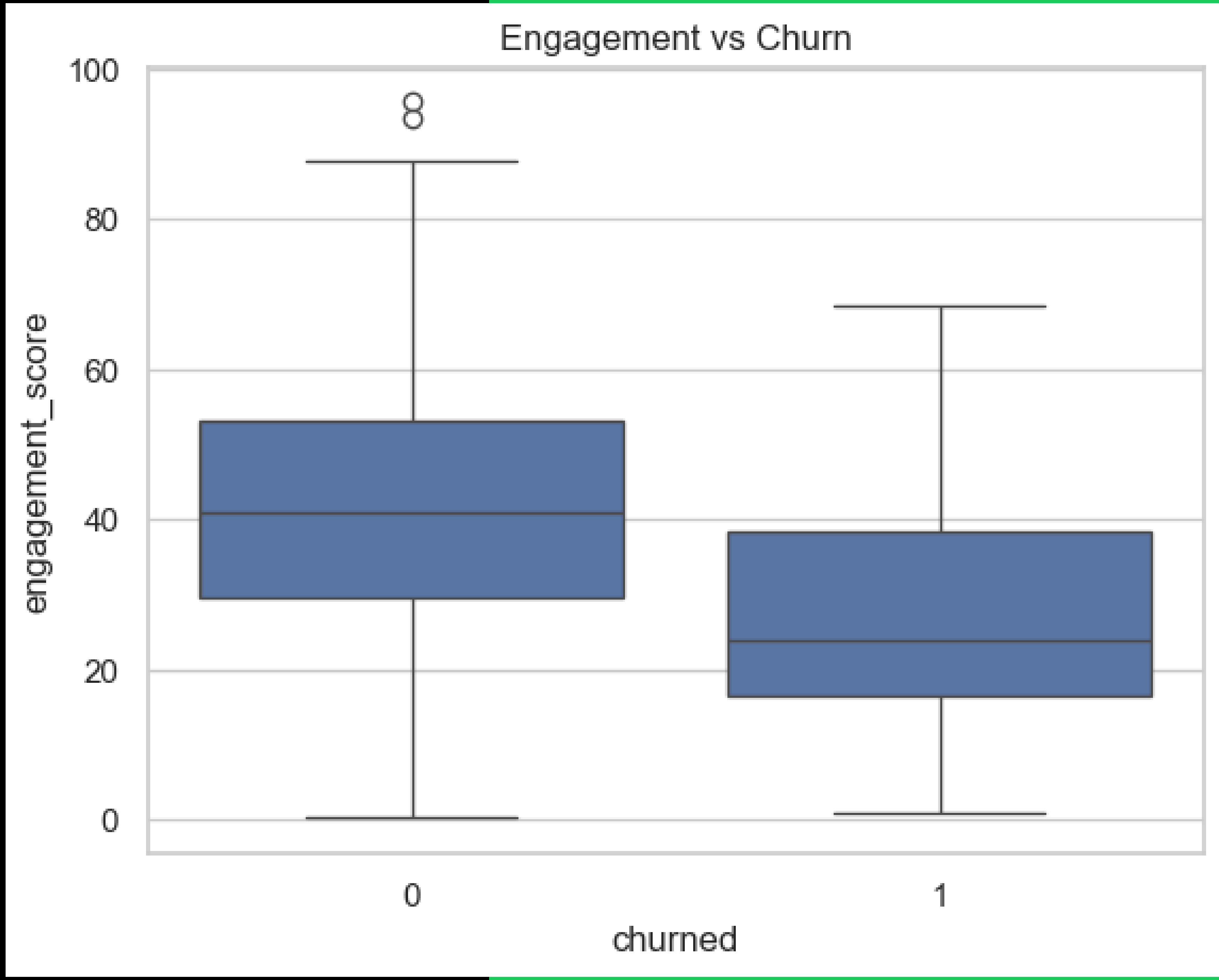
MODEL	ACCURACY	RECALL (CHURN)	PRECISION	F1 (CHURN)
Logistic Regression	0.75	0.73	0.41	0.52
Random Forest (Tuned)	0.78	0.38	0.41	0.39
Original RF	0.81	0.24	0.50	0.33

E. COMPARATIVE ANALYSIS

- Users with fewer listening minutes, longer inactivity, and fewer playlists are more likely to churn.
- These behavioral metrics are clear indicators of disengagement.

F. ENGAGEMENT SCORE VS. CHURN

- Engagement Score combines recency, frequency, skips, and time spent.
- Churners show a significantly lower average engagement score.
- Effective as a high-level indicator for early churn detection.



INFERENCE

Logistic regression, with class balancing, yielded the best recall, essential for churn prevention.

Behavioral features (like avg_daily_minutes, last_login) were much more predictive than demographic ones.

Engagement score is a powerful, interpretable summary feature.

SUGGESTIONS TO

REDUCE CHURN

- Target users with declining engagement scores through retention emails or personalized offers.
- Provide incentives for playlist creation and more session time.
- Detect and act on users who haven’t logged in for a long time.
- Improve user experience for those frequently skipping or opening tickets.

End Slide