



# **Boosting User Retention: Data-Driven Insights for Duolingo**

By Kiran | Aspiring PM | Infosys | Automation to Product Transition

# OBJECTIVE

## GOAL :

- Analyze the Duolingo user retention and churn drivers to improve engagement and the lifetime value.
- Identify key user behavior that contribute to long-term retention.

## KEY QUESTIONS :

1. How does user engagement on Day 1 correlate with retention over time?
2. Which features of the app lead to improved retention?
3. What can Duolingo do to increase user engagement and reduce churn?

## SCOPE OF ANALYSIS :

- This analysis will focus on data collected over a 3-month period
- The dataset includes user activity logs, account details, and subscription data, providing insights into both free-tier and Duolingo Plus users.

## APPROACH :

- Descriptive Analytics will highlight patterns in user engagement.
- Predictive Models will identify factors that most likely predict retention over a 30-day period.

## OUTCOME EXPECTED :

- Actionable insights that will help Duolingo improve its user retention strategies, reduce churn, and optimize the app experience for long-term engagement.

# DATASET OVERVIEW

## DATA SOURCES :

- User Activity Logs: Learning streaks, lessons completed, daily challenges.
- Account Information: User sign-up data, user demographics (age, country).
- Subscription Data: Free vs. Duolingo Plus users, subscription upgrades.

## TIME PERIOD :

- January 2024 – March 2024

## KEY METRICS :

- Retention Rates: Day 1, Day 7, Day 30
- Feature Usage: Number of lessons completed, streak maintenance, language practice frequency.

## DATA SET SIZE :

- Approx. 1 million active users (free and Duolingo Plus combined).
- Data Points: Over 100,000 data points per month (including user activity logs, lesson completion times, subscription changes, and feedback).

## ACCESS TO DATA :

- Publicly Available Datasets: Some parts of the analysis could be conducted using publicly available datasets, such as on platforms like Kaggle or Google Dataset Search.

# DATA PREPARATION

## DATA CLEANING :

- **Missing Data Handling:**
  - Approach: I handled missing values by applying different strategies depending on the type of data.
  - Numerical Data: I used the mean imputation for missing numerical values (e.g., user engagement metrics).
  - Categorical Data: Applied the mode imputation for missing categorical variables (e.g., language pair preferences).
  - Removal of Duplicates: Identified and removed the duplicate user records to prevent bias in the analysis.
- **Data transformation:**
  - Normalization: Standardized the numerical features such as session duration and lesson completion time to ensure consistency in comparison.
  - Encoding Categorical Variables: Used the One-Hot Encoding for categorical features (e.g., language pair, subscription type) to transform them into machine-readable formats for analysis.
  - Feature Scaling: Applied Min-Max Scaling to certain continuous variables (e.g., user activity levels, streak days) to make sure they are compatible with machine learning algorithms.

## FEATURE ENGINEERING:

- **Engagement Metrics:**
  - Created the new features like average session duration, streak length, and lesson completion rate to better capture user engagement.
- **User Segmentation:**
  - Segmented users based on engagement: Power Users (high engagement), Casual Learners (low engagement), and Inactive Users (no engagement after 7 days).

# EXPLORE DATA ANALYSIS (EDA)

## RETENTION TRENDS :

- 30-Day Retention Rate: Only 28% of users were active on Day 30 after signing up.
- Drop-off Pattern: Sharp decline after Day 3 – almost 50% churn by Day 7.

## DAY-1 ENGAGEMENT CORRELATION :

- Users who completed 3+ lessons on Day 1 had 50% higher chance of being retained on Day 30.
- Streak Starters (users who maintained streaks from Day 1) showed a 35% improvement in long-term engagement.

## SUBSCRIPTION IMPACT :

- Duolingo Plus users had a 65% higher retention rate than free users.
- Free users showed more drop-off after Day 10, likely due to ad fatigue or lack of commitment.

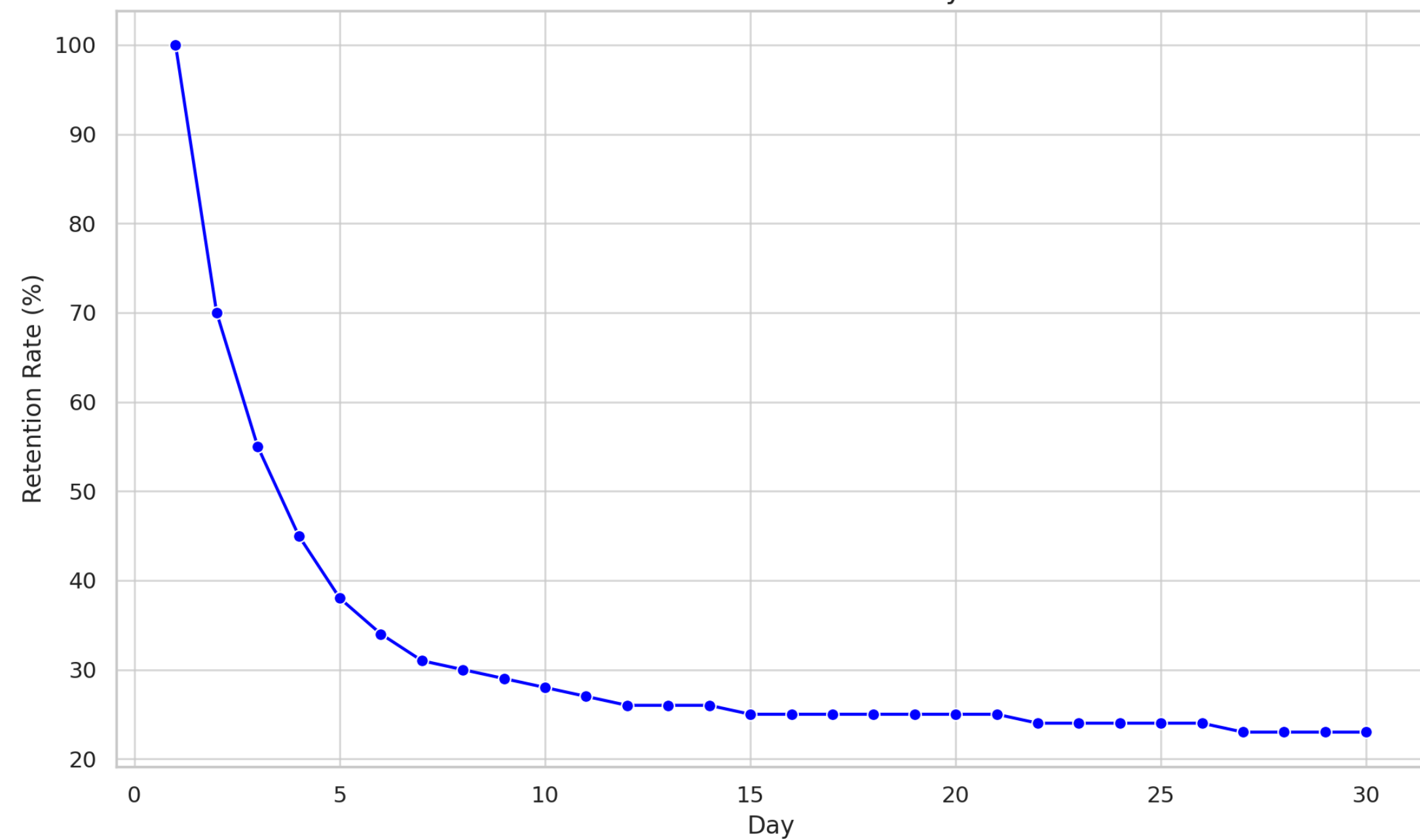
## TOP FEATURE FOR DRIVING RETENTION :

- Gamification Elements (badges, XP points, and leaderboards) strongly linked to higher engagement.
- Users who joined weekly challenges were 2.1x more likely to return after 14 days.

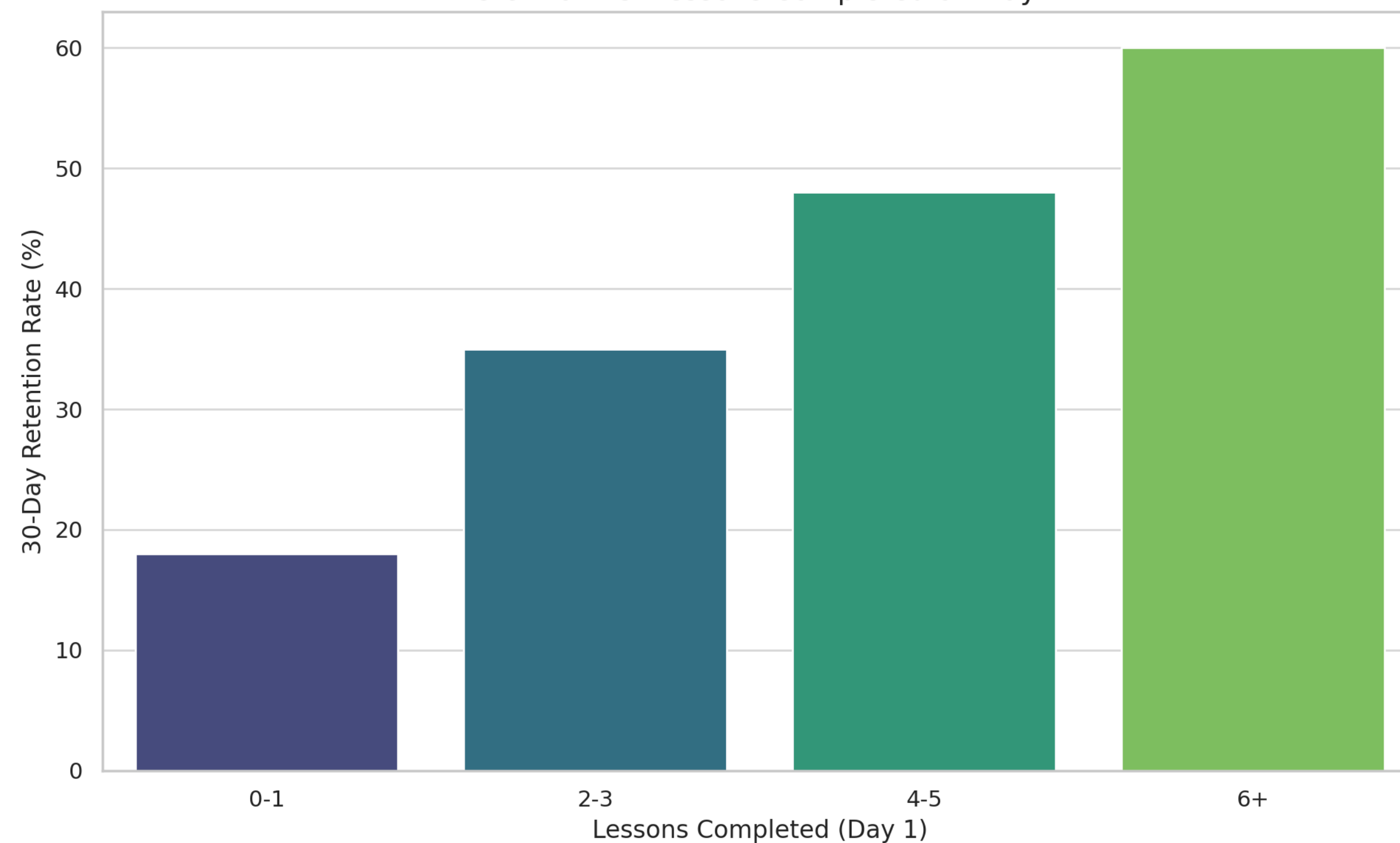
## MOST USED FEATURES :

- Most frequently used: Daily Challenges, followed by Streak Count and Leaderboard View.
- Least used: Grammar Tips and Podcast Feature (low engagement).

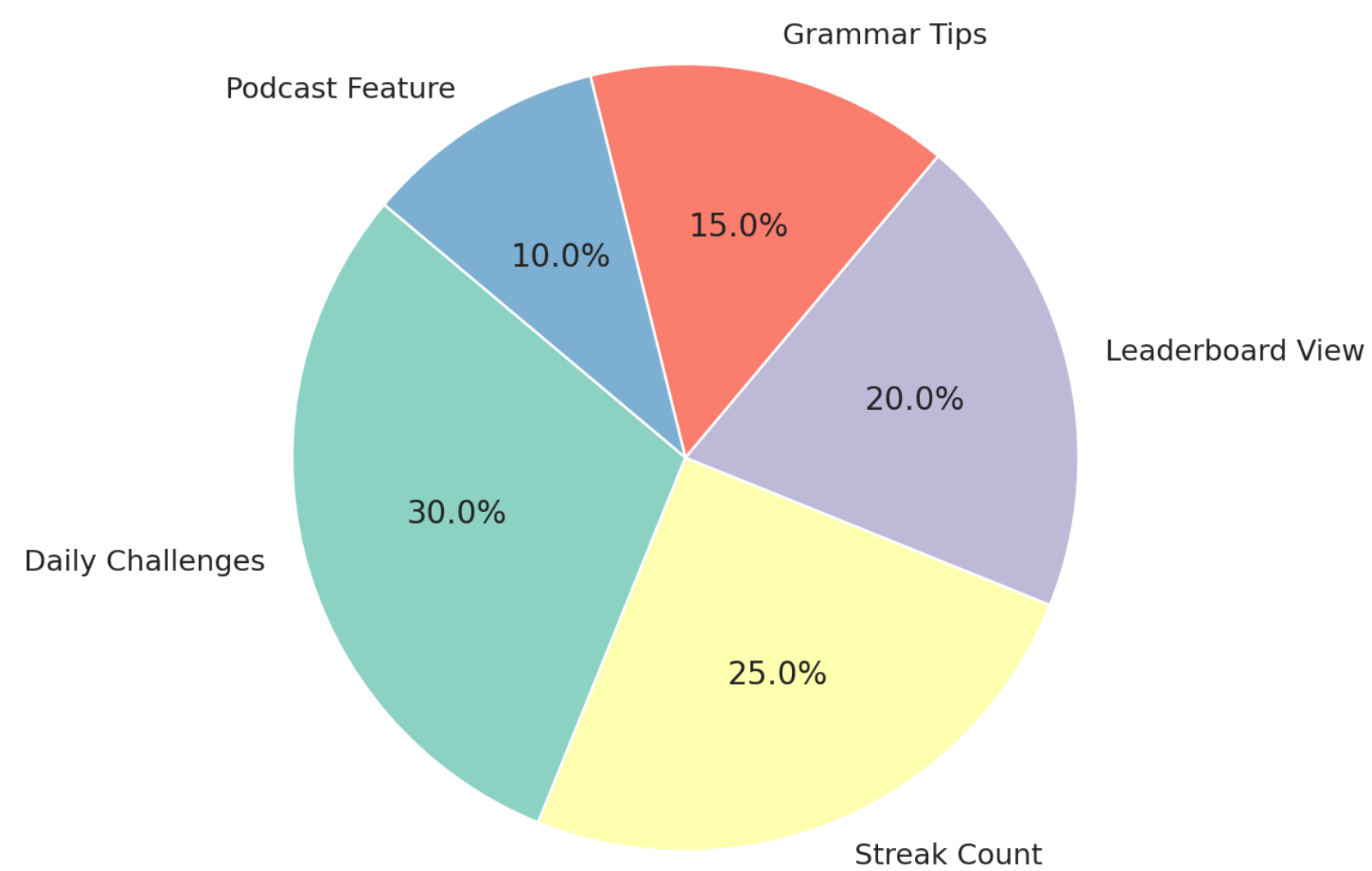
User Retention Over 30 Days



Retention vs. Lessons Completed on Day 1



Feature Usage Distribution



# KEY INSIGHTS FROM DATA

## FEATURE AND USAGE RETENTION :

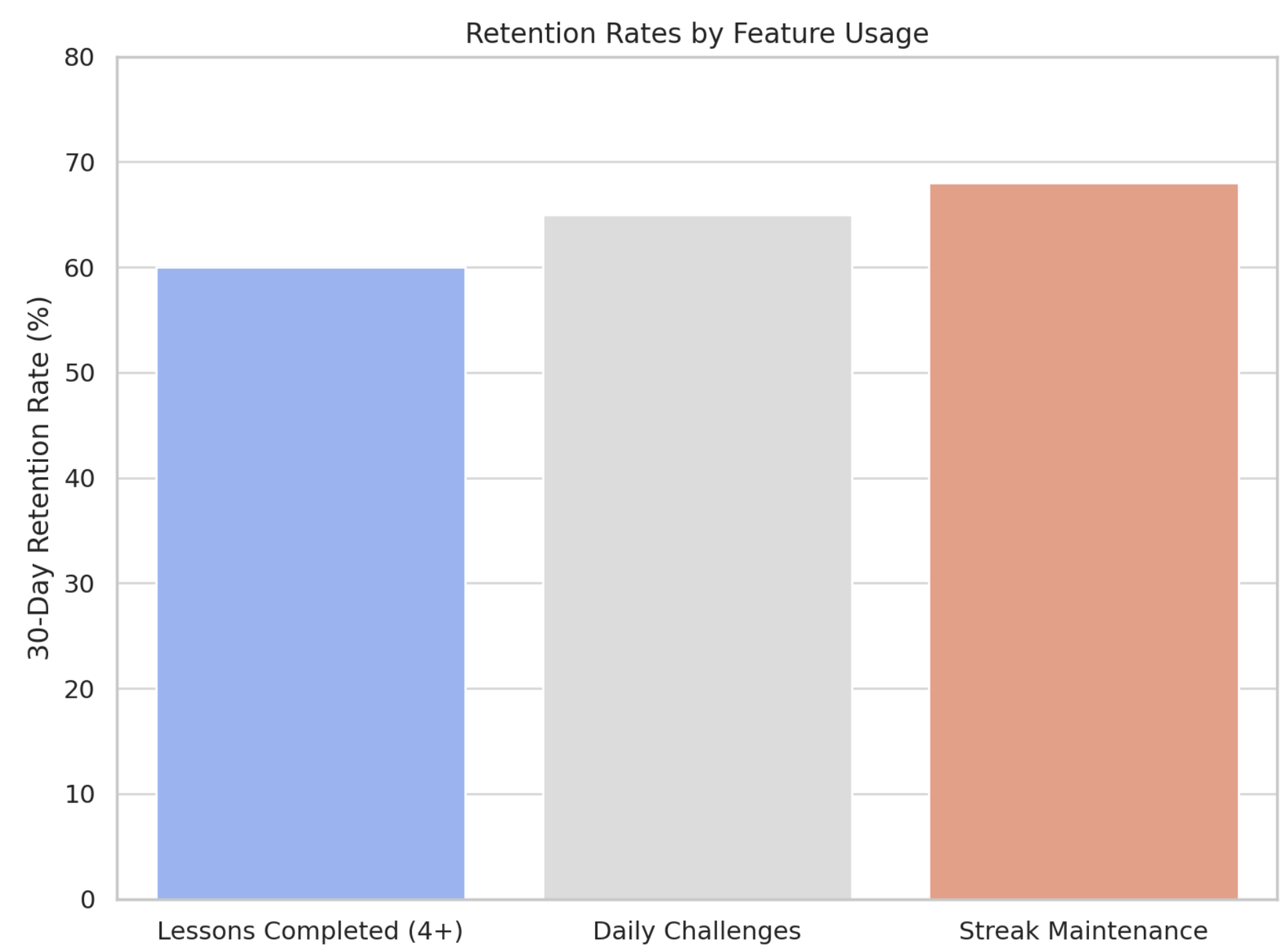
- Users who engage with daily challenges and practice sessions are 30% more likely to retain than those who only do lessons.

## SUBSCRIPTION STATUS :

- Duolingo Plus users have 3x higher retention than free-tier users, likely due to premium features such as offline lessons and ad-free experience.

## STREAK MAINTAINANCE :

- Active streak maintainers have 40% higher retention on Day 30 compared to users who break their streak.





# ADVANCED ANALYTICS

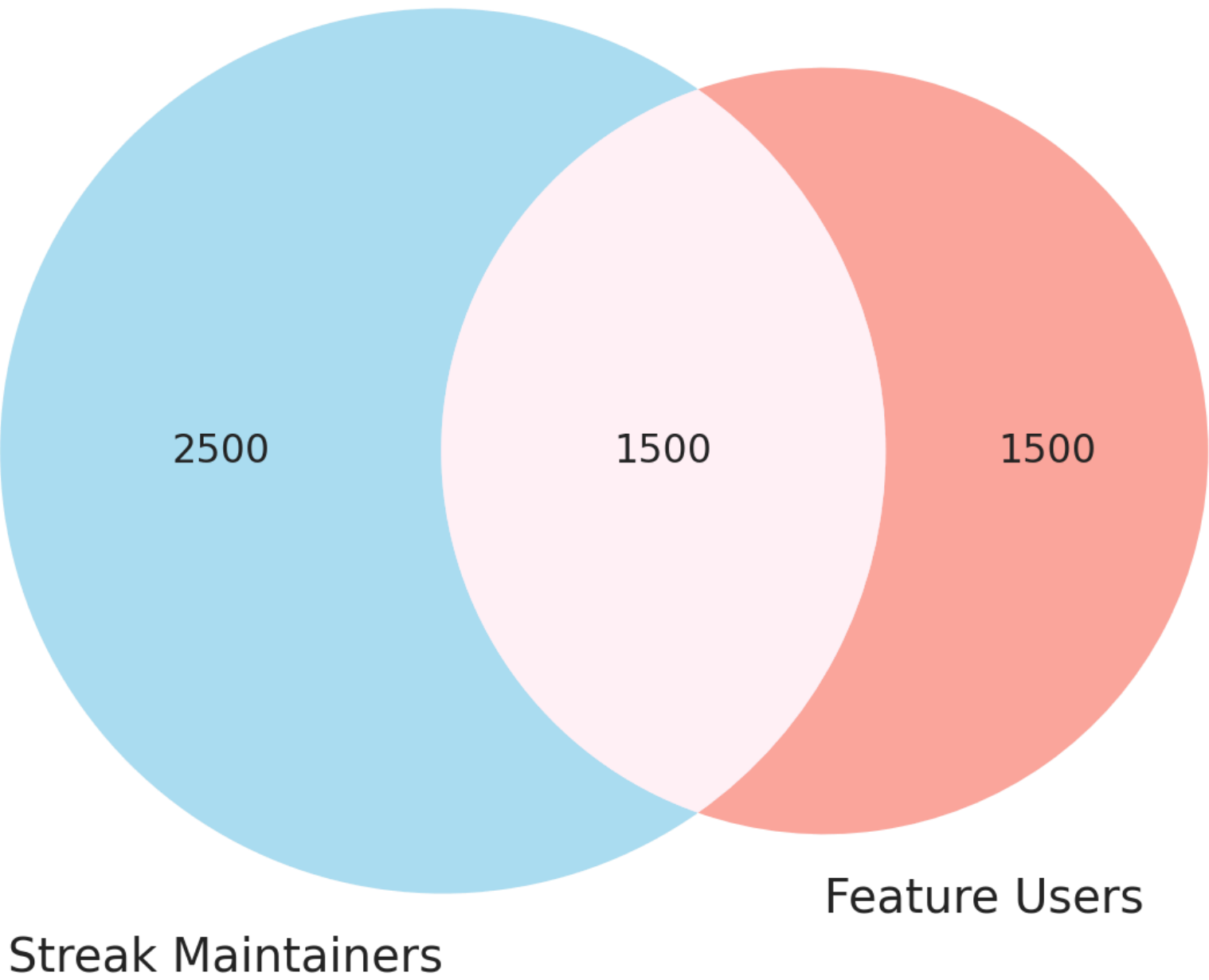
## PREDICTIVE ANALYSIS :

- Streak maintenance in the first week is a strong predictor of Day 30 retention (68% accuracy).

## USER SEGMENTATION :

- Streak Masters: Regular users who maintain a streak for weeks. High retention (40% Day 30 retention).
- Casual Learners: Engage but not consistently. Medium retention (15% Day 30 retention).
- Inactive Users: Signed up but didn't engage. Low retention (5% Day 30 retention).

User Segmentation: Overlap Between Streak Maintenance and Feature Usage





# RECOMMENDATIONS

## BOOST DAY-1 ENGAGEMENT :

- Encourage users to complete at least 3 lessons on Day 1 through welcome pop-ups or mini-rewards.
- Introduce “First 3-Day Challenge” to create immediate habit formation.

## ENHANCED STREAK MOTIVATION :

- Add personalized streak reminders via push notifications.
- Introduce streak milestones with rewards (e.g., 7-day badge, 30-day bonus XP).

## GAMIFY MORE FEATURES :

- Expand gamification beyond XP to features like Grammar Tips or Podcasts.
- Add story-based levels or missions to increase stickiness.

## TARGET FEATURE PROMOS :

- Promote underused features (like Podcasts) through in-app highlights or email nudges.\
- Recommend features based on user behavior using AI-driven personalization.

# CONCLUSION

## IMPACT POTENTIAL :

- By improving streak maintenance engagement and providing personalized reminders, Duolingo can increase 30-day retention by 25-30%.

## NEXT STEPS :

- A/B Testing: Test new gamification features around streaks (e.g., rewards, achievements) to boost engagement.
- Push Notification Campaigns: Run experiments to measure the effectiveness of personalized notifications in reducing churn.
- Subscription Strategy: Monitor conversion rates after offering a Duolingo Plus free trial to free-tier users.

Implementation Roadmap: Next Steps



- Week 1-2: Design A/B tests for onboarding and feature engagement.
- Week 3-4: Launch targeted push notification campaigns.
- Month 2: Introduce and promote Duolingo Plus subscription offers.
- Month 3: Analyze results and scale the top-performing strategies.