Duolingo User Churn Analysis Report

Introduction

This project explores key factors contributing to user churn on Duolingo, aiming to provide actionable recommendations to improve retention. The analysis addresses two primary questions:

1. What causes users to churn?
2. What actionable strategies can reduce churn?

To answer these questions, the project followed a structured approach involving exploratory analysis, correlation analysis, predictive modeling, and advanced modeling.

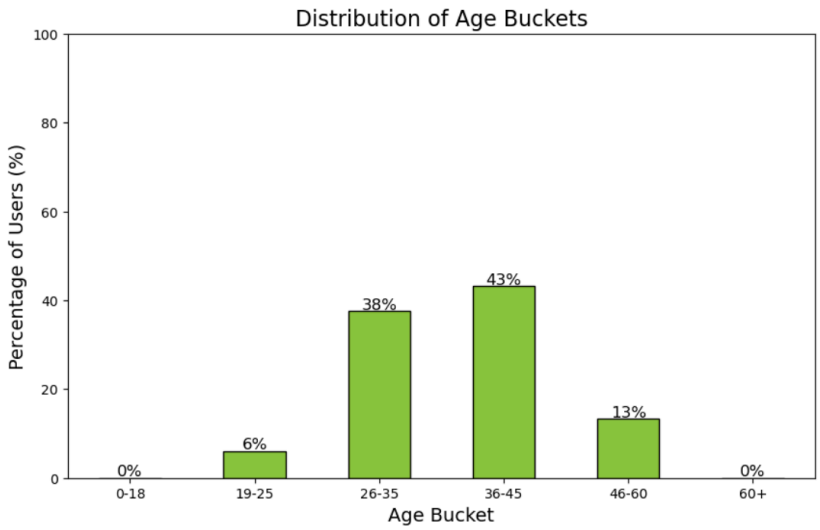
Approach:

1. **Exploratory Analysis**: Analyzed demographic and behavioral data to uncover general trends and patterns.
2. **Correlation with Churn**: Identified variables strongly correlated with churn to understand characteristics most associated with user retention.
3. **Predictive Modeling**: Developed a logistic regression model to quantify the influence of different factors on churn.
4. **Advanced Modeling with XGBoost**: Leveraged XGBoost and created a model with grid search to uncover patterns and interactions among features.

Exploratory Analysis

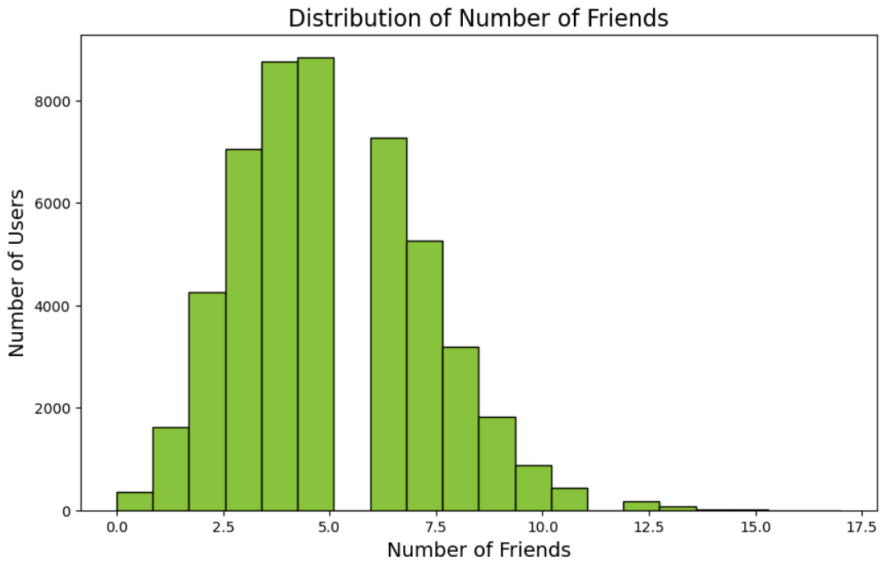
Initial exploratory analysis revealed key insights into user demographics and behaviors:

- **Demographics**:
    - 81% of users are aged 26-45, highlighting a younger to middle-aged user base.

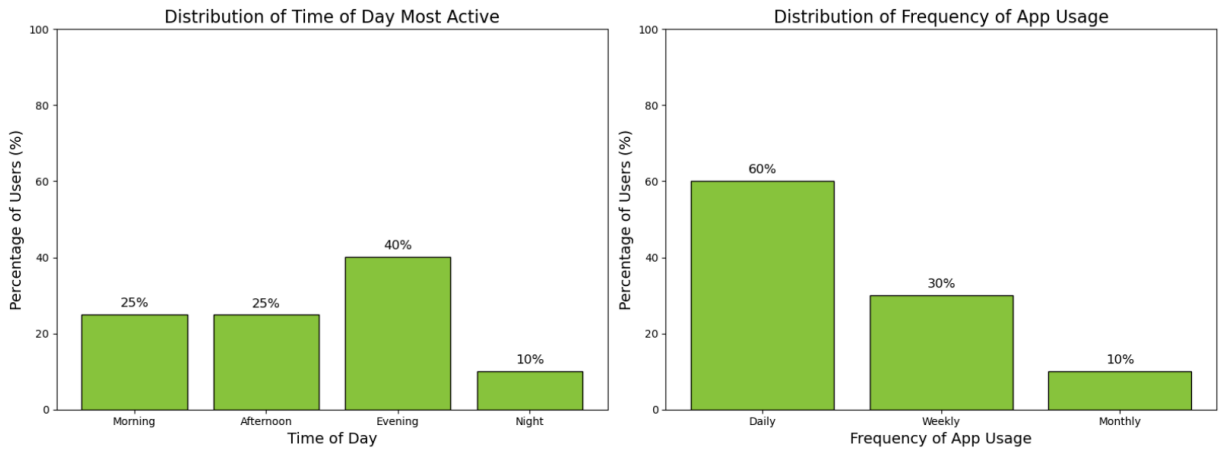


- Gender distribution is balanced with 48% males and 47% females, with no significant bias.

- **Behavioral Trends**:
  - Most users have 2-10 friends added on Duolingo, but the distribution is right-skewed.



  - Peak activity occurs in the evening, likely due to users having more free time.
  - 60% of users are daily active, and 30% are weekly active users.



- **Retention Observations**:
    - Users with doctoral degrees have a 0% churn rate, despite representing 8% of the user base.
    - Churn rates are consistent across different friend count categories, suggesting that friend count is not a significant predictor of churn.

Correlation with Churn

- **Age**: Older users are more likely to churn, possibly due to changing priorities or reduced engagement with the platform.
- **Average Sessions Per Week**: Users with higher weekly session counts are less likely to churn, highlighting the importance of consistent engagement.

Logistic Regression Analysis

To predict churn and identify key factors, I built a logistic regression model.

Key Predictors:

1. **Age**: Each additional year increases the odds of churn by **27.3%**.
2. **Average Sessions Per Week**: Each additional session reduces the odds of churn by **40.2%**.
3. **Number of Languages Mastered**: Each additional mastered language reduces churn odds by **51.1%**.
4. **Subscription Type**: Premium subscribers are **35.8%** less likely to churn.

Model Performance:

- **Variance Inflation Factor (VIF)**: No multicollinearity was detected among the features.

- **AUC-ROC**: The model achieved a score of **92%**, indicating strong performance.

Feature Selection and Regularization:

Lasso regularization was applied to select the most influential features. Grid search was used to determine the best parameters for ross-validation ensured the model generalized well to unseen data.

XGBoost Modeling

An XGBoost model was implemented to capture complex interactions among features and improve predictive accuracy.

- **Model 1 (Default Parameters)**:
  - Accuracy: **0.8953**
  - AUC-ROC: **0.9310**
- **Model 2 (with Cross Validation)**:
  - Accuracy: **0.8915**
  - AUC-ROC: **0.9294**
- **Model 3 (with Cross Validation and Grid Search)**:
  - Accuracy: **0.8944**
  - AUC-ROC: **0.9315**

We will use Model 3, XGBoost with grid search and cross validation as it produces a model that ensures that hyperparameters are optimized for performance to generalize well on unseen data with slightly higher ROC-AUC and accuracy.

Most Influential Features:

1. **Age:** Younger users of Duolingo exhibit higher engagement levels and lower churn likelihood, while older users may face barriers like time constraints or lower tech familiarity,
2. **Average Sessions Per Week**: Higher values indicate better engagement, reducing churn likelihood.
3. **Total Lessons Completed**: Reflects user progress and investment in the platform.
4. **Last Active Days Ago**: Higher inactivity days signal disengagement and increased churn risk.

Recommendations

The model's output enables actionable strategies for customer retention:

1. **Target High-Risk Customers**: Identify users with churn probabilities over 70% and implement retention campaigns such as discounts or personalized support.

2. **Monitor Medium-Risk Customers**: Focus on users with churn probabilities between 40-70% by observing their behavior and encouraging engagement.
3. **Engage Low-Risk Customers**: Retain loyalty among users with low churn probabilities through rewards or premium features.

Considerations for Future Datasets

1. **New Features**: Retrain the model if new features are introduced to fully leverage their predictive power.
2. **Distributional Changes**: Validate that the new data distribution matches the original training set to maintain model performance.
3. **Hyperparameter Tuning**: Re-tune hyperparameters if significant differences exist between the original and new datasets.

Conclusion

This project provided insights into the key drivers of churn and actionable recommendations for user retention. The XGBoost model, supported by exploratory analysis and logistic regression, serves as a robust tool for predicting churn and guiding retention strategies. With these findings, Duolingo can better focus on fostering consistent engagement and catering to high-risk users to reduce churn effectively.