Retention in Gamified Learning: Analyzing Duolingo User Engagement with Machine Learning

Kaz Fukuhara, Xinman (Yoyo) Liu, Xinyu (Teah) Shi
{kazf28, xinman, teah2001}@stanford.edu

Predicting

Language learning apps like Duolingo serve 37.2 million daily active users, yet many students drop off rapidly, making user retention a critical challenge. Using the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM) dataset including 5,520 users' first 30 days, we examined how engagement patterns predict retention by processing behavioral features (correctness, response times, engagement frequency) through Softmax, Neural Network, Random Forest, PCA, and Agglomerative Hierarchical Clustering models. We found that engagement frequency and consistency significantly outweighed performance accuracy in predicting user retention. PCA revealed four fundamental engagement dimensions, leading to four distinct user profiles that inform targeted retention strategies.

Data

We used the 2018 Duolingo SLAM dataset containing 5,520 users' first 30 days of language learning activity. The raw data consisted of 2,243,983 token-level rows (French-English, Spanish-English, and English-Spanish) with features capturing linguistic information, user responses, exercise formats, and timing data. Each token included correctness (whether answered correctly), response time, and metadata about the session and user. We aggregated this to user-level with an 80-20 train/test split using stratified sampling. The target variable is the churn time category (1: 0-6 days, 2: 7-13 days, 3: 14-20 days, 4: 21+ days).

Features

We engineered 17 features across five categories:

- correctness metrics (block/reverse tap/translate/listen average correctness)
- response time metrics (average times, variance)
- user ability score (combined measure of correctness and normalized response time)
- activity metrics (average blocks per session/day, average sessions per day, total blocks and sessions completed)
- consistency measures (variances in session length, blocks per session, blocks per day, sessions per day)

Features were standardized using log transformation, QuantileTransformer, and MinMaxScaler.

Methods

Softmax Classifier: Converts raw output scores into probability distribution over the four churn categories. Serves as baseline model with linear decision boundaries and clear feature weight interpretability.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K=4} e^{z_j}}$$

Neural Network: Three-layer architecture with ReLU activation functions, batch normalization to stabilize learning, and RMSprop optimization which performs better than Adam with our small sample size. We implemented early stopping to prevent overfitting once validation loss began increasing.

$$f(X) = \sigma(W_3 \cdot \mathrm{ReLU}(W_2 \cdot \mathrm{ReLU}(W_1 \cdot X + b_1) + b_2) + b_3)$$

Random Forest: Ensemble of decision trees (optimized to 50 estimators, max depth of 10) using information gain for split selection. Hyperparameters were optimized through grid search with 10-fold cross-validation across 576 parameter combinations.

$$IG(D,\ a) \ = \ H(D) \ - \ \frac{|D_j|}{|D|} H(D_j)$$

$$H(D) \ = \ -\, p_i \log_2(p_i)$$

PCA: PCA reduces dimensions to 4 principal components (selected using modified Kaiser criterion with eigenvalues > 1.5) capturing 60% of variance.

$$\Sigma = \frac{1}{m} X^T X$$

$$[U, S, V] = \mathrm{SVD}(\Sigma)$$

$$Z = XU_{reduce}$$

Agglomerative Hierarchical Clustering: We conducted hierarchical clustering on the PCA dimensions with Ward's linkage to minimize within-cluster variance, which creates more balanced, compact clusters particularly useful for behavioral segmentation.

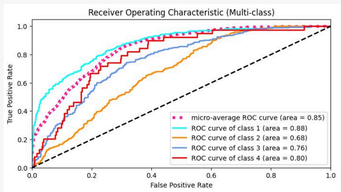
$$d(C_i, C_j) = \sqrt{\frac{2|C_i||C_j|}{|C_i| + |C_j|}} ||m_i - m_j||_2$$

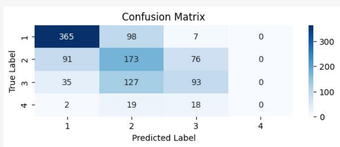
Results

	Training Error	Test Error	AUC
Softmax	0.442	0.451	0.83
Neural Network	0.427	0.446	0.85
Random Forest	0	0.429	0.83
Neural Network with SMOTE	0.524	0.525	0.83
Random Forest with SMOTE	0	0.476	0.81

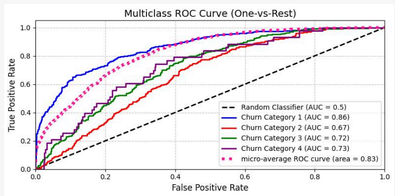
We trained our supervised models on 4,416 user examples (80%) and tested on 1,104 examples (20%), with the Neural Network achieving the best performance (AUC=0.85) while Random Forest and Softmax also performed strongly (AUC=0.83). Note RF with 0 train error before test error which suggests initial overfitting.

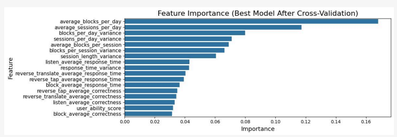
Neural Network





Random Forest

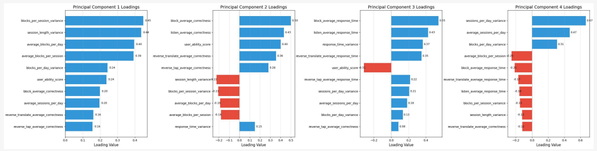




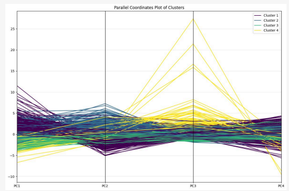
Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE was applied to address class imbalance, but the modest decline in AUC performance with synthetic data suggests that our initial models may have effectively captured the signal within the original dataset.

PCA and Hierarchical Clustering



Principal Components: 1) Engagement Variability & Intensity; 2) Learning Performance vs. Structure; 3) Response Time & Learning Efficiency; 4) Learning Schedule Variability



User Profile

- Persistent Challengers (32.1%): High engagement despite struggling with performance
- Efficient Achievers (25.9%): High performance with moderate effort
- Casual Explorers (36.1%): Minimal commitment across most metrics
- Deep Thinkers (5.9%): Deliberate approach with longer response times

Discussion

Our analysis demonstrates that behavioral engagement patterns—specifically frequency and consistency of interaction—possess substantially greater predictive power than performance accuracy metrics, challenging purely competence-based models of learning (Chrzan et al., 2023). We found that early disengagement patterns where users drop out at approximately 1 week were most confidently identified, aligning with Duolingo's discovery that user retention increases significantly with 10-day streaks (Mazal, 2023). The four distinct user profiles we uncovered through PCA and clustering provide empirical support for engagement-centered theories focused on habit formation rather than performance improvement. These findings were unexpected but compelling, suggesting that language learning platforms should prioritize establishing behavioral routines through targeted strategies: detailed explanations for "Persistent Challengers," competitive elements for "Efficient Achievers," bite-sized goals for "Casual Explorers," and untimed exercises for "Deep Thinkers." Our results indicate that retention is driven more by establishing behavioral habits than by promoting competence-based engagement.

Future Work

Future work could implement sequential modeling approaches like LSTMs to capture temporal patterns and identify critical transition points for each user profile. Additionally, A/B testing protocols for targeted retention interventions could also be developed to measure their effectiveness in real-world scenarios. Finally, incorporating demographic and motivational data would provide deeper context for behavioral differences, enabling more nuanced retention strategies across user segments.

References

Chrzan, M., Li, T. J., & Sparks, A. (2023). Moving from Users to Learners: Analyzing Duolingo User Behavior Data to Predict User Persistence Using Semi-Supervised Learning. Stanford University.
Gustafson, E. (2023, February 17). Meaningful metrics: How data sharpened the focus of product teams. Duolingo Blog. https://blog.duolingo.com/growth-model-duolingo/
Hwang, H.-B., Coss, M. D., Loewen, S., & Tagarelli, K. M. (2024). Acceptance and engagement patterns of mobile-assisted language learning among non-conventional adult L2 learners: A survival analysis. Studies in Second Language Acquisition, 46(4), 969–995. https://doi.org/10.1017/S0272263124000354
Isbell, D., Rawal, H., Oh, R., & Loewen, S. (2017). Narrative Perspectives on Self-Directed Foreign Language Learning in a Computer- and Mobile-Assisted Language Learning Context. Languages, 2(2), 4. https://doi.org/10.3390/languages2020004
Li, M., & Wang, T. (2024). Optimizing learning return on investment: Identifying learning strategies based on user behavior characteristic in language learning applications. Education and Information Technologies, 29(6), 6651–6681. https://doi.org/10.1007/s10639-023-12078-9
Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. ReCALL, 31(3), 293–311. https://doi.org/10.1017/S0958344019000065
Mazal, J. (2023, February 28). How Duolingo reignited user growth. Lenny's Newsletter. https://www.lennysnewsletter.com/p/how-duolingo-reignited-user-growth
Nithya, S., & Umarani, S. (2022). MOOC Dropout Prediction using FIAR-ANN Model based on Learner Behavioral Features. International Journal of Advanced Computer Science and Applications, 13(9), Article 9. https://doi.org/10.14569/IJACSA.2022.0130972
Settles, B. (2018). Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM) (Version 4.0) [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/8SWHNO
Ünal, E., & Güngör, F. (2021). The continuance intention of users toward mobile assisted language learning: The case of DuoLingo. Asian Journal of Distance Education, 16(2). https://www.semanticscholar.org/paper/The-continuance-intention-of-users-toward-mobile-of-%C3%9Cnal-G%C3%BCng%C3%B6r/b16e1ffe08806a9df2899b1fcedab4f98bb0d154
Yang, Z. (2022). FCM Clustering on Interaction Pattern Analysis of Chinese Language Learner Behavior. Computational Intelligence and Neuroscience, 2022, 8256646. https://doi.org/10.1155/2022/8256646