Moving from Users to Learners: Analyzing Duolingo User Behavior Data to Predict User Persistence Using Semi-Supervised Learning

Michael Chrzan, Tracy Junfan Li, Alexa Sparks
Stanford University

Abstract

This research study focuses on analyzing user behavior data from Duolingo, a language learning app. The primary objective is to predict user churn using semi-supervised learning techniques. By examining factors that contribute to user disengagement within the first 30 days, the study aims to provide actionable insights for improving user retention. The methodology involves employing Principal Component Analysis (PCA), Agglomerative Hierarchical Clustering, and XGBoost machine learning models. Notably, our findings highlight that predictors indicating user behaviors on the app, particularly the quality of the time spent on the app, serve as key predictors of when a user is likely to discontinue using Duolingo in these first 30 days. We hope these insights can inform strategies for enhancing user retention on Duolingo and other apps like it.

Introduction

Second Language Acquisition (SLA), the study of individuals or groups learning a second language, encompasses a variety of learning modalities, ranging from informal environments - such as a child acquiring a new language by immersion without formal instruction - to formal education settings. Technological advancements have spawned numerous sub-disciplines within SLA. Mobile-assisted language learning (MALL), for example, focuses on understanding the impact of mobile devices on language acquisition with an emphasis on gamification, personalized learning experiences, and the potential to transcend traditional barriers of time and space for greater accessibility.[1] Duolingo is one of many language learning apps that fall into this category, representing a significant case study for MALL's effectiveness in engaging and retaining users.

Duolingo's analysis of resurrected users, those who return to the app after a period of inactivity, emphasizes the complexity of user engagement patterns on the platform. They found that resurrected users often exhibit lower retention rates upon their return, underscoring a need for strategies that not only re-engage users but also retain them.[2] This study aims to build upon Duolingo's findings by focusing on user engagement during their first 30 days using the app. We

[1] Loewen et al., "Mobile-Assisted Language Learning."
[2] Walsh, "Back from the Brink."

seek to identify factors that contribute to disengagement, defined as the user's drop day, and provide suggestions to minimize the necessity for re-engagement efforts and improve overall user retention.

Our approach was informed by Duolingo's 2018 Shared Task on Second Language Acquisition Modeling (SLAM) competition's insights and the predictive framework developed by Barbaro et al. for predicting user engagement in mobile applications. The SLA competition demonstrated the efficacy of gradient boosted decision trees for modeling complex, non-linear relationships in language learning data.[3] Barbaro et al.'s research introduced a framework for quantifying engagement metrics, employing Agglomerative Hierarchical Clustering to segment users into meaningful groups, and using XGBoost models to predict user drop day.[4] For this study, we used Principal Component Analysis (PCA) to reduce dimensionality before clustering user data.

To predict each user's drop day, four XGBoost machine learning models were trained: 1) on the raw aggregated dataset, 2) on the four PCA dimensions used for clustering users, 3) only on the cluster assignments, and 4) on both the PCA dimensions and cluster assignments. The models were evaluated using R^2 (explaining variance in drop day) and root mean squared error (RMSE). The models trained on just the PCA dimensions (2) and the PCA dimensions plus clusters (4) performed best, explaining around 15% more variance than the raw data model (1) and 65% more than the clusters-only model (3), with the lowest RMSE of around 14 days on average. Results on the test set showed these top models were robust against overfitting the training data. The PCA + Clusters model (4) performed slightly better than PCA-only (2) during cross-validation. Examining feature importances from the PCA + Clusters model (4) revealed that dimensions capturing how users spent their time, rather than outcomes like accuracy, were most predictive of drop day. This insight emphasizes the importance of considering user behavior and engagement patterns beyond mere performance metrics, suggesting a nuanced approach to enhancing persistence in language learning through mobile platforms.

Research Question

Which specific user characteristics, if any, have significant influence on user retention on Duolingo?

Methods

Dataset

We use a subset of Duolingo's 2018 SLAM competition dataset that includes 2,528 English speaking users learning Spanish during the first 30 days of using Duolingo.

[3] Dureddy, Larionov, and Qian, "Second Language Acquisition Modeling - Duolingo."
[4] Barbaro et al., "Modelling and Predicting User Engagement in Mobile Applications."

Data Preprocessing

Data Cleaning

In the data documentation there was a note that said negative values for time were caused due to client-side browser bugs on the web.[5] We also noticed in our exploratory analysis that the maximum value for time was excessively high at 38.12 hours. We only included users who spent between 0 and 3 hours on the app in our analysis.

Feature Selection

To operationalize and quantify user engagement, we engineered a comprehensive set of features based on the raw data available for each user, including countries, days, sessions, exercise formats, and time. In alignment with the suggestion of McFarland and McFarland, we created features that describe the recency, frequency, and value (in terms of accuracy) of users in order to help create our clusters.[6] These derived features capture key aspects of engagement at multiple granularities: daily averages, session-level averages, overall totals, and contextualized measurements.

At the daily level, we computed *avg_total_accuracy* (the mean accuracy across all sessions for a given day), *avg_session_time*, *avg_exercises_per_session*, and *avg_num_sessions_in_day*. These average metrics provide a snapshot of the user's typical daily interaction intensity and language learning progress within the app.

We also calculated session-level average accuracy metrics contextualized by exercise format (*avg_exercise_accuracy, avg_reverse_trans_accuracy, avg_reverse_tap_accuracy, avg_listen_accuracy) and session type (avg_session_accuracy, avg_test_accuracy, avg_lesson_accuracy, avg_practice_accuracy*). Differentiating accuracy by context enables assessment of whether users exhibit varying levels of success that could influence their motivation to persist.

To quantify the totality of a user's engagement, we derived several aggregate metrics: *country_count* (number of distinct countries the app was accessed from), *total_time* (the cumulative time spent on the app), *total_exercises*, *total_sessions*, and totals specific to session types (e.g., *total_test_sessions*). These totals serve as measures of the user's overall commitment, effort expenditure, and exposure to varied learning modalities over their lifetime using the app.

Finally, we operationalized indicators of disengagement behavior through *days_skipped_user* (the maximum number of consecutive days without app activity) and *last_day_accuracy*. Collectively, this feature set aims to comprehensively characterize multi-faceted user engagement patterns and motivational factors across temporal segments.

[5] "2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)."
[6] McFarland and McFarland, "Big Data and the Danger of Being Precisely Inaccurate."

Response Variable

To analyze user attrition on Duolingo, we calculated the *drop_day* variable, which represents each user's last active day on the app. This was determined by identifying the maximum value of the *days* variable, which indicates the number of days since each user started learning a language on the platform. We rounded the days variable down using the float method to ensure we got the day the user dropped and not the day after if their final usage was later in their last day.

Exploratory Analysis

Figure 1 presents a visual representation of user attrition over the first 30 days on Duolingo. It is evident from the graph that there is a consistent decrease in the number of users returning each day. Notably, the number of returning users diminishes to zero by day 28, indicating that all users stopped using the app before reaching the 30 day mark. This trend suggests that user engagement wanes during the first 30 days; our analysis seeks to explore the underlying patterns of this decline.

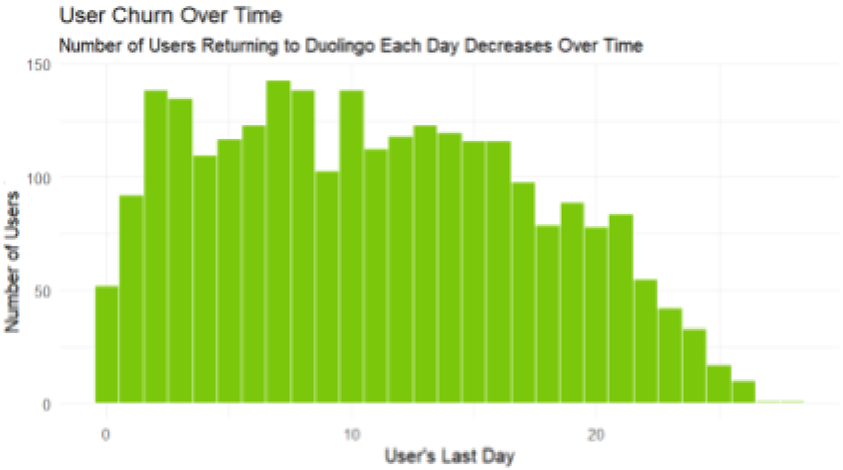


Figure 1 - Number of users each day of the 30 days in the dataset.
Note: the last user of the dataset leaves on day 28.

Analytical Methods

To begin our analysis of the derived user characteristics, we first reduce the dimensionality of the data using Principal Component Analysis (PCA). We then use the resulting components to define clusters of users using Agglomerative Hierarchical Clustering (HC). After completing our principal component interpretation, we examine how each cluster is described by the components (or dimensions) to build profiles of the types of users present in the dataset and describe their characteristics. Finally, we generate four predictive models using XGBoost trained on different combinations of the variables and their reductions and compare the results to determine how to best predict a new user’s final day on the app.

Results

PCA Results

Our principal component analysis reveals that between 58.4% and 69.1% of the variance in our latent variables can be explained with the first 3 to 5 dimensions. The next step is deciding the number of dimensions to keep for our analysis. A common rule-of-thumb for this is choosing an "elbow" in the scree plot, or an inflection point where the return of variance for adding a new dimension starts to dwindle.

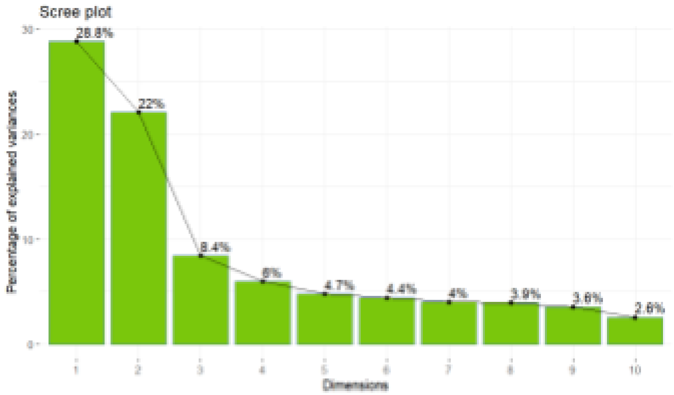


Figure 2 - Scree plot from the PCA on the latent variables derived from the original data

Examining the scree plot for our data in Figure 2, we see that any of the first 3-5 dimensions could serve as a viable "elbow" given the slow descent over these three dimensions. Due to the lack of a clear elbow, we turn to Silva et. al.'s empirical assessment of criteria for choosing the number of dimensions in a principal component analysis, where they show that choosing dimensions based on the percentage of variance explained (with a target of 70%) is a suitable rule for estimating the number of PCA axes.[7] We balance their result with a desire to keep our interpretation manageable in the next steps of our analysis and decide to include the first 4 dimensions, resulting in an explanation of 65.2% of the variance.

With the number of dimensions to retain decided, we turn to examining those dimensions in order to interpret, as much as possible, what characterizes those four dimensions.

[7] Silva et al., "Criteria for Choosing the Number of Dimensions in a Principal Component Analysis."

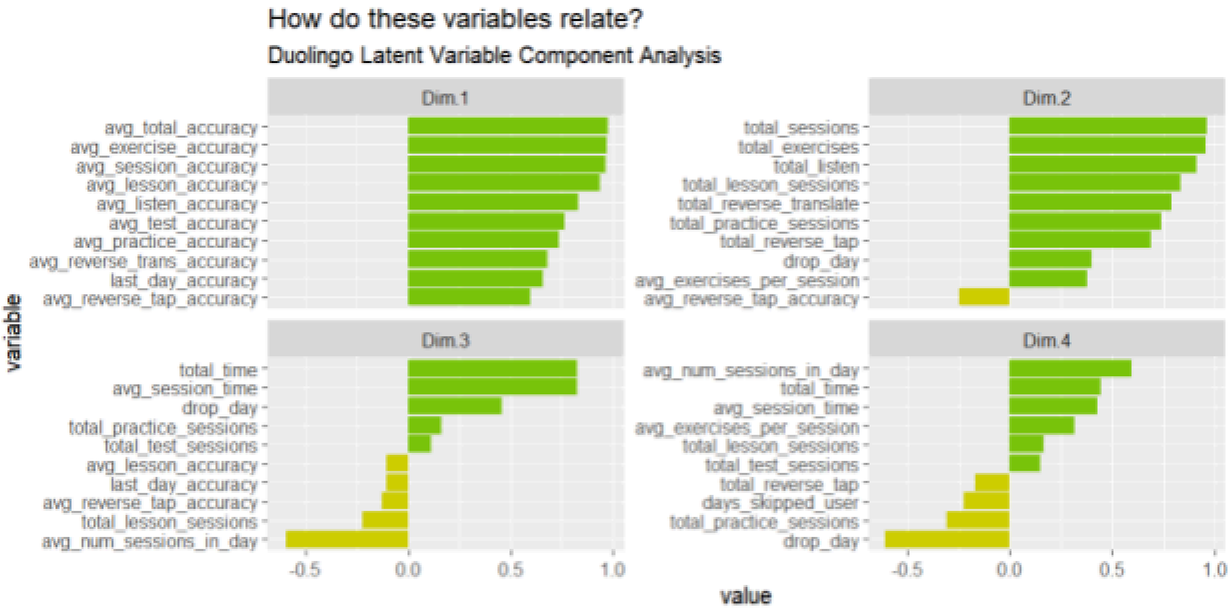


Figure 3 - Component Loadings from PCA

Dimension 1. Accuracy

The first of these dimensions is perhaps the most straightforward to interpret. As we can see in Figure 3, each variable with high importance in this dimension is a variable measuring some aspect of user accuracy. Users high in this dimension are ones who are more accurate. We note that all the loadings are also positive, indicating a strong correlation between these different measures of accuracy and supporting our decision to use a dimension reduction technique to analyze the data.

Dimension 2. Investment

The second dimension also has a fairly straightforward interpretation. In this dimension, we see numerous variables that indicate activity or engagement levels. For that reason we describe this dimension as a measure of investment from the users, where users who score highly on this dimension are ones doing many more activities in the app, with potentially varying accuracies captured in the first dimension.

Dimension 3. Rushedness

In analyzing the third dimension, we acknowledge that there is more room for interpretation. Here we see that the variables associated with time have high positive loadings on this dimension while variables associated with accuracy have negative loadings. Given the discrepancy in the loadings for different types of activities as well (positive for tests and practices, but negative for lessons), we characterize this dimension as measuring how fast users rush through the app. Users high in this dimension are ones who rush to tests and practices, maybe mistaking past learning of the language for mastery early on, but struggle to recall the language enough to be accurate.

Dimension 4. Burnout

Dimension four, like dimension three, is up to interpretation. For our analysis, we note that this dimension also seems to capture some measure of time spent on different activities. Variables like *avg_num_sessions_in_day*, *total_time*, *avg_session_time*, and *avg_exercise_session_time* have high positive loadings, suggesting this dimension represents the time or duration aspect of various activities or sessions. However, the negative loadings on variables such as *days_skipped_user* and - in particular for how negative it is - *drop_day* indicate that users high in this dimension are users who use the app a lot and don't miss days, but then struggle to keep up with the app and have much earlier drop days than other users.

Cluster Results

Now equipped with clear ideas about how we are measuring users, we move to examine particular groups of users as measured on these dimensions to identify types of early Duolingo users represented in this dataset.

To create these groupings, we turn to Agglomerative Hierarchical Clustering. We chose this clustering method to align with the framework laid out in Barbaro et. al., where they analyzed similar metrics for an industrial dataset built in the area of waste recycling. However, here we decide to use a different distance metric for our analysis; namely we use Ward's distance.

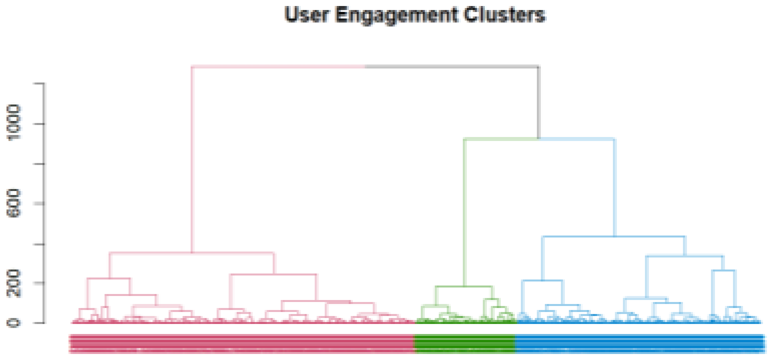


Figure 4 - Dendrogram Resulting from Agglomerative Hierarchical Clustering on PCA Dimensions

Ward's distance is a popular choice for hierarchical clustering because it aims to minimize the total within-cluster variance. This method tends to create clusters that are relatively compact and spherical in shape, which is desirable when dealing with data where clusters are expected to have a more homogeneous structure. Here we expect such structure due to the type of data collected, where all the users are new to learning the language on the app. In the instances where that assumption doesn't hold, Ward's method is also relatively robust to the presence of outliers and tends to group outliers into their own small clusters, rather than distorting the overall clustering structure.

Perhaps the most useful affordance to our analysis this clustering method provides with this linkage is that, as we can see in Figure 4, the resulting dendrogram for this clustering method reveals that there are multiple levels available to define clusters. This helps us examine meaningful numbers of clusters to have and gives some insight into the data's underlying patterns. Here we choose to create three clusters. We do so for the following reasons:

1. Clear separation: The dendrogram shows a distinct separation of the data points into three main branches or clusters at a relatively large height on the y-axis. This large vertical distance suggests that the three clusters are well-separated and distinct from each other.
2. Cluster compactness: Within each of the three main branches, the data points are merged at relatively small heights, indicating that the points within each cluster are closely related and compact. This compactness is a desirable property achieved by our choice of Ward's distance.
3. Interpretability: Having three distinct clusters aids in interpreting and understanding the underlying patterns or characteristics of the data.
4. Balancing complexity and granularity: Choosing three clusters strikes a balance between oversimplifying the data by having too few clusters and introducing unnecessary complexity by having too many clusters. Three clusters can provide a reasonable level of granularity while still maintaining interpretability and a clear separation of the data.

Building User Profiles

Now we move to the final step of our user analysis, building user profiles. This step is an amalgamation of the previous, where we analyze how the clusters we created with hierarchical clustering project onto the dimensions yielded by our PCA model and use our interpretations of those dimensions to describe the clusters of users. This summary is provided in Figure 5, which we use to construct our analysis of the three clusters.

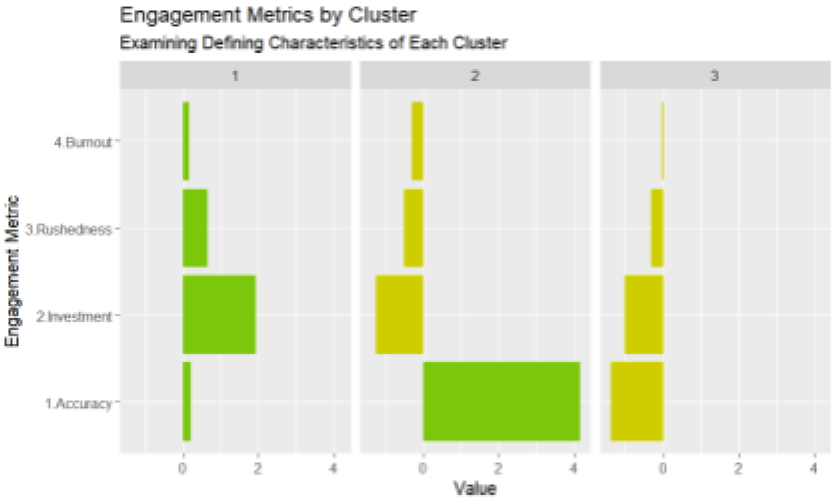


Figure 5 - PCA Component Loadings for Each Cluster

Cluster 1 Profile

We see cluster 1 as defined by their investment. We see this cluster as the traditional learners; these are users who are invested in their learning and, even though they aren't right most of the time since they are new to the language, they're putting in the time to get better. We arrive at this due to the large loading in the second dimension, the investment dimension, as well as positive loadings on all three other dimensions, but a noticeably small loading on the burnout dimension.

Cluster 2 Profile

This cluster is heavily defined by their accuracy. We see this cluster as containing users who may already be familiar with the language, since they have such high accuracy but low time invested. Their lower values on the rushedness and burnout dimensions also lead us to believe that these users are rushing to tests and practices and are more likely to skip days and drop out earlier.

Cluster 3 Profile

Lastly, cluster 3 is what we affectionately dub “the bad place”. These are users who we believe use Duolingo more like a game than a language learning service. While marked by similar usage patterns as cluster 2 (in terms of their loadings’ direction and magnitude for investment, rushedness, and burnout) these users also have the unfortunate reality of having very low accuracy. These are users Duolingo may want to particularly target more to convince them to migrate into usage patterns more closely aligned with cluster 1 due to their presumed lack of prior knowledge of the language.

Prediction Model Results

Now that we have full profiles for the types of users, we move toward answering our research question - which of these user characteristics, if any, have significant influence on user retention on Duolingo?

To answer this question, we trained an XGBoost model to predict each user's last day on the app, which we called their “drop day". We chose XGBoost, for its ability to handle the complexity of the data and pick up small signals in data, like that provided by the drop_day variable. We trained four such models for this purpose, all using the XGBoost algorithm:

1. A model trained on the raw data from our aggregation of the original dataset, which was the only model with the specific variables we generated.
2. A model trained on the four PCA dimensions we used to define our clusters.
3. A model trained only on the clusters.
4. A model trained on both the PCA dimensions and the clusters

Each of these models was trained on an 80/20 train/test split of the users and was trained with a 10-fold cross-validation method. The results of these models are summarized in Tables 1A and 1B and Figure 6 below.

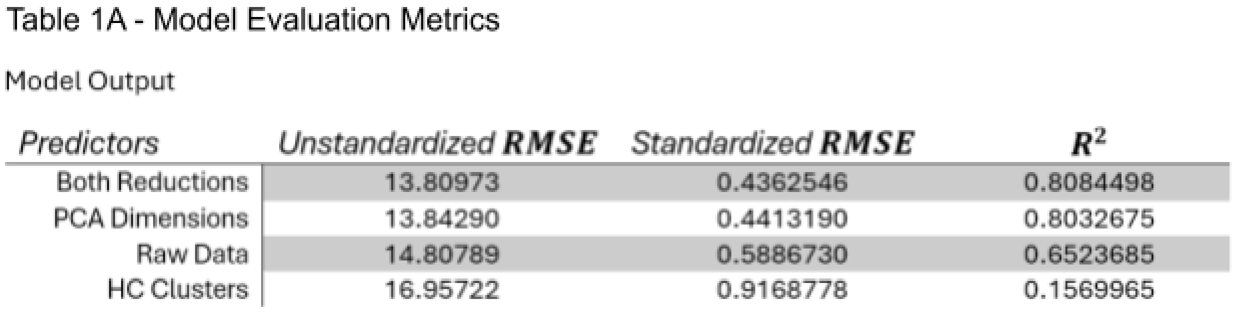


Table 1A - Model Evaluation Metrics

Model Output

Predictors	*Unstandardized* ***RMSE***	*Standardized* ***RMSE***	$\boldsymbol{R^2}$
Both Reductions	13.80973	0.4362546	0.8084498
PCA Dimensions	13.84290	0.4413190	0.8032675
Raw Data	14.80789	0.5886730	0.6523685
HC Clusters	16.95722	0.9168778	0.1569965

In Table 1A, we see that the model's trained on solely the PCA dimensions and the model trained on those dimensions with the cluster assignment meaningfully outperform the other two models, accounting for 15% more variance than the raw data and 65% more variance than the clusters alone. We also see that this corresponds to these two models having the lowest average RMSE, an average of the measure of the prediction's deviation from the true result across all samples in the cross-validation. Here, the RMSE is unstandardized so that the units are directly interpretable as the average number of days of error in our predictions, which for the best model sits at around 14 days.

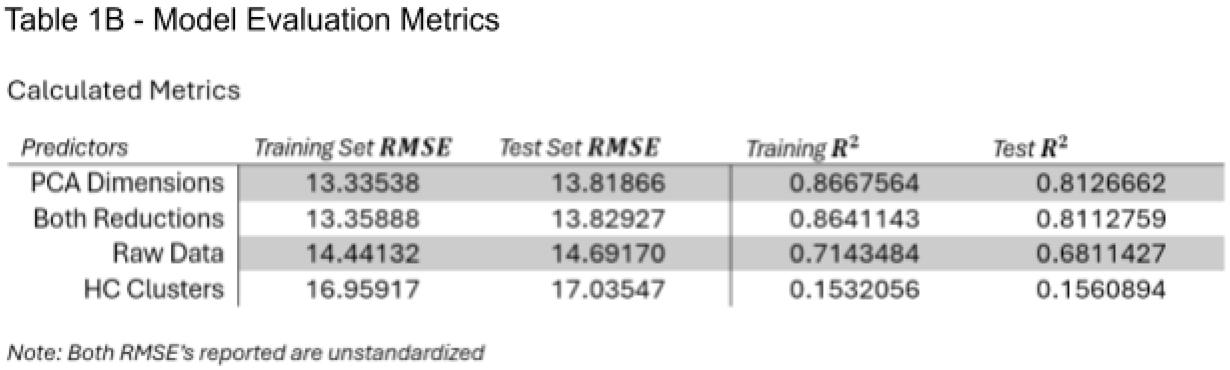


Table 1B - Model Evaluation Metrics

Calculated Metrics

Predictors	*Training Set* ***RMSE***	*Test Set* ***RMSE***	*Training* $\boldsymbol{R^2}$	*Test* $\boldsymbol{R^2}$
PCA Dimensions	13.33538	13.81866	0.8667564	0.8126662
Both Reductions	13.35888	13.82927	0.8641143	0.8112759
Raw Data	14.44132	14.69170	0.7143484	0.6811427
HC Clusters	16.95917	17.03547	0.1532056	0.1560894

Note: Both RMSE's reported are unstandardized

In Table 1B, we see similar patterns emerge for these metrics measured on the training and test set predictions. We note the slight decrease in RMSE for the training data predictions, but a similar RMSE to the model's average for the training data provided in Table 1A, leading us to believe the model is robust against overfitting the training data. We also note a similar phenomenon in the R^2 metric for the testing data, where the training R^2 greatly increases for the prediction data but then matches or slightly improves upon the model average for the test R^2. These results are further bolstered by our visualization of the spread of the predictions provided in the appendix in Figure A4.

Another key difference is that, here, the model with only PCA Dimensions as predictors comes out slightly above the model trained on both reduction methods. Since these changes are so slight and the models trained on the PCA versus PCA + Clusters perform so similarly, yet on the cross-validated sets of training data the PCA + Clusters performs better and includes slightly more information by way of having another predictor, we decide to examine the importance provided by the PCA + Cluster model.

In Figure 6 we see the importance of each PCA dimension for the model's predictions. This figure allows us to examine how all of the work thus far comes together to understand the relationship between the individual measures we have for users and user churn.

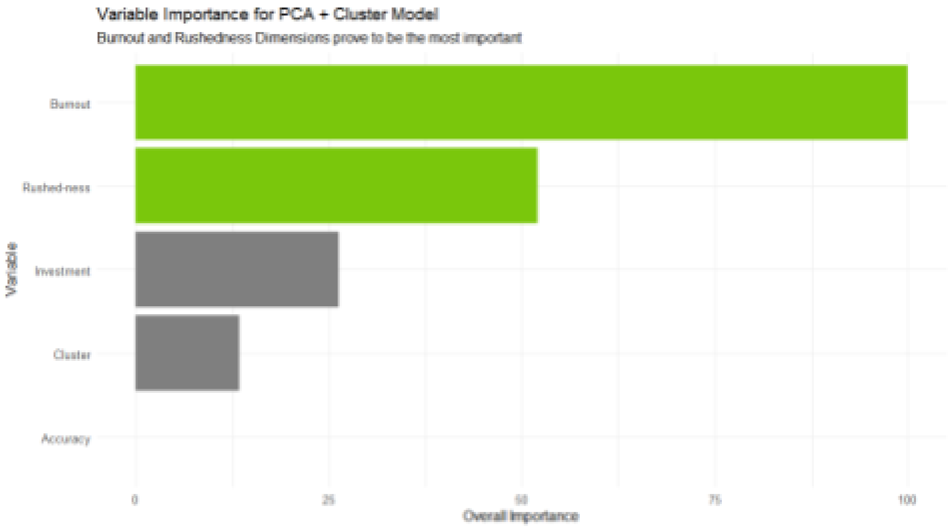


Figure 6 - Variable Importance for drop_day predictions

In this figure we discover something potentially counterintuitive. The most important aspects when predicting the last day a user will use Duolingo are not the time they are investing into the app or the accuracy they achieve on the app (both of which are things heavily reinforced in Duolingo through features such as leaderboards). Here we see that the dimensions measuring *how users spend their time* - rather than the *outcomes* of that time, like accuracy - are more useful in predicting when a user will stop using the app.

Discussion

Our study presents meaningful takeaways for Duolingo's user retention strategy. As of 2023, Duolingo tracks ~27 million Daily Active Users (DAU) and ~89 million Monthly Active Users (MAU), with a paid subscription penetration rate of ~7.8% for MAUs.[8] Duolingo has indicated that business strategy for 2024 will entail 1) teaching more effectively and; 2) relentless focus on user retention.[9] They intend to do this in various ways such as experimenting with session types, engagement styles, and application layout (including through AI-driven personalization tools) and by offering new user experience packages (including the freemium, super, and max application versions) while building off their family subscription packages, and capitalizing on Duolingo's paid English proficiency certification offering.[10]

Duolingo can improve the way it experiments with personalization of learning and engagement by sending more personalized push notifications to users. For example, they could send notifications that account for whether a skipped day on the app is indicative of possible burnout, low commitment and low achievement, or no real risk/threat to long-term engagement. For low commitment and low achievement individuals, this segment may not have a low subscription

[8] Duolingo, "Duolingo Shareholder Letter Q4 FY2023."
[9] Ibid.
[10] Ibid.

conversion rate. They may, however, be a valuable group for promoting the application further among their social networks. For users exhibiting potential signs of burnout, it may be worthy to experiment with in-app rewards to re-engage these users, rather than use negative reinforcement push notifications or email nudges.

To improve the accuracy of business growth forecasting, Duolingo can also build on our model to forecast total future daily active users and active monthly users with important user data. While the RMSE of our model has room for improvement, we are confident that model accuracy can be made more precise through the addition of other relevant user metrics, which we did not have access to for this study. Improved business growth forecasting will lead Duolingo to more accurately allocate capital across new research projects, major investments in marketing initiatives, partnerships, and new application feature offerings.

Future research directions may include testing other clustering methods for user classification, and performing the same research on different user datasets. This includes: other language courses, other time ranges of data, particular user demographics (i.e. country of residence, gender, age, in-app purchasers). We believe it would also be worthwhile to explore conversion rates between our identified user profiles with monetization metrics, be it a purchase of super, max, and/or family-tier subscriptions (and purchase renewals), other one-time in-app purchases, or special products like Duolingo's more expensive English language certification test (or future certification tests in other languages, like Spanish). We believe these insights will be particularly valuable for informing how Duolingo will transform the freemium SAL digital learning market, but also find grounding among the more established formal SAL digital learning market.

Conclusion

Using a subset of Duolingo's 2018 SLAM competition dataset with 2,528 English speaking users learning Spanish during the first 30 days, we trained four XGBoost machine learning models to predict a user's drop day. Rather than focus on predicting word token accuracy using the data like other researchers have in the past, we wanted to understand how user behavior predicted their retention, as retention is a key business goal for Duolingo. Our best performing model used both principal components and hierarchical clustering assignments to predict user drop day. This model had a root mean squared error (RMSE) of predicting drop day within a 14-day range (+/- 7 days) accuracy.

The most important contributing factors for our models were how users spent their time, rather than their outcomes (i.e. accuracy). We found that 1) total time spent; 2) total sessions completed during a single app engagement; 3) skipped days; 4) types of sessions completed (i.e. practice, lesson, test) contributed most to predicting user drop day. Our study defines "drop day" as the last active use day during the 30-days of data which we use.

Our study results are relevant for Duolingo to forecast user growth over time, but also to improve user targeting by using the user profiles we identify to provide the appropriate types of motivational nudges. Our study results are also relevant for Duolingo to identify user conversion trends in future studies, where user profiles may be studied in the context of how it affects in-app purchases, or subscription behavior. Lastly, our study adds to the literature of user engagement in online environments by adding validation evidence to the process laid out in Barbado et. al. by applying it to a new, broader context.

Future research directions could include repeating the model training with added user metrics (i.e. demographic data) or model training on different subsets of data, such as exploring model performance with other second language learners and over longer time ranges than the 30 days provided here.

Acknowledgements

The authors would like to thank Duolingo for making their dataset publicly available for research purposes. This allows the work of the budding field of Education Data Science to have access to meaningful data on learning that would otherwise be privatized and inscrutable. We would also like to thank our support team of instructors, classmates, and cohort mates for helping us set up the dataset from what Duolingo had publicly available and for their feedback throughout the process of designing our study and analyzing our results. Particular thanks to Dr. Sanne Smith for her instruction on Machine Learning methods this quarter that made this research possible.

References

"2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)." Accessed March 16, 2024. http://sharedtask.duolingo.com.

Barbaro, Eduardo, Eoin Martino Grua, Ivano Malavolta, Mirjana Stercevic, Esther Weusthof, and Jeroen van den Hoven. "Modelling and Predicting User Engagement in Mobile Applications." *Data Science* 3, no. 2 (January 1, 2020): 61–77. https://doi.org/10.3233/DS-190027.

Duolingo. "Duolingo Shareholder Letter Q4 FY2023". Duolingo, January 15, 2024. https://investors.duolingo.com/static-files/06cda5ae-c66f-4d99-82ec-da764ecb1034

Dureddy, Hima, George Larionov, and Xin Qian. "Second Language Acquisition Modeling - Duolingo," n.d.

Loewen, Shawn, Dustin Crowther, Daniel R. Isbell, Kathy Minhye Kim, Jeffrey Maloney, Zachary F. Miller, and Hima Rawal. "Mobile-Assisted Language Learning: A Duolingo Case Study." *ReCALL* 31, no. 3 (September 2019): 293–311. https://doi.org/10.1017/S0958344019000065.

McFarland, Daniel A, and H Richard McFarland. "Big Data and the Danger of Being Precisely Inaccurate," 2015. https://doi.org/10.1177/2053951715602495.

Silva, Renata B., Daniel de Oliveira, Davi P. Santos, Lucio F. D. Santos, Rodrigo E. Wilson, and Marcos Bedo. "Criteria for Choosing the Number of Dimensions in a Principal Component Analysis: An Empirical Assessment." In *Anais Do Simpósio Brasileiro de Banco de Dados (SBBD)*, 145–50. SBC, 2020. https://doi.org/10.5753/sbbd.2020.13632.

Walsh, Connor. "Back from the Brink: What Duolingo Learned about Its Resurrected Users."

Duolingo Blog, August 30, 2017. https://blog.duolingo.com/back-from-the-brink-what-duolingo-learned-about-its-resurrected-users/.

Appendix

Code: All code for this project can be found at the following GitHub repository mlchrzan/Duolingo-User-Interaction-Modeling-Project (github.com).

Data: Here is the data we worked with (Formatted from released data here, code for conversion here). The original dataset (here) contains 2 million words (tokens) from answers submitted by more than 6,000 students over the course of their first 30 days of using Duolingo for 3 language tracks: English, Spanish and French. This study only focuses on one language pair, Spanish and English. Anonymized user IDs and time data is provided, which allows for sequential/ longitudinal analysis.

Figure A1 - Examining Cluster Placement on PCA Dimensions 1 and 2

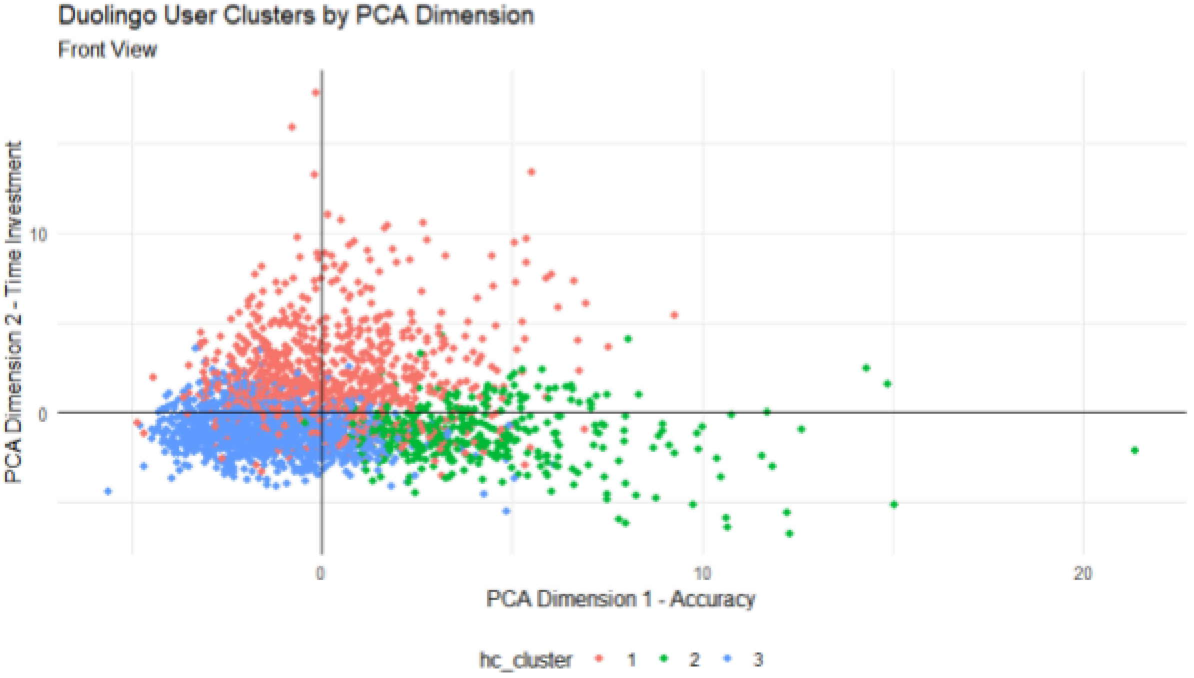


Figure A2 - Examining Cluster Positions on PCA Dimensions 3 and 2

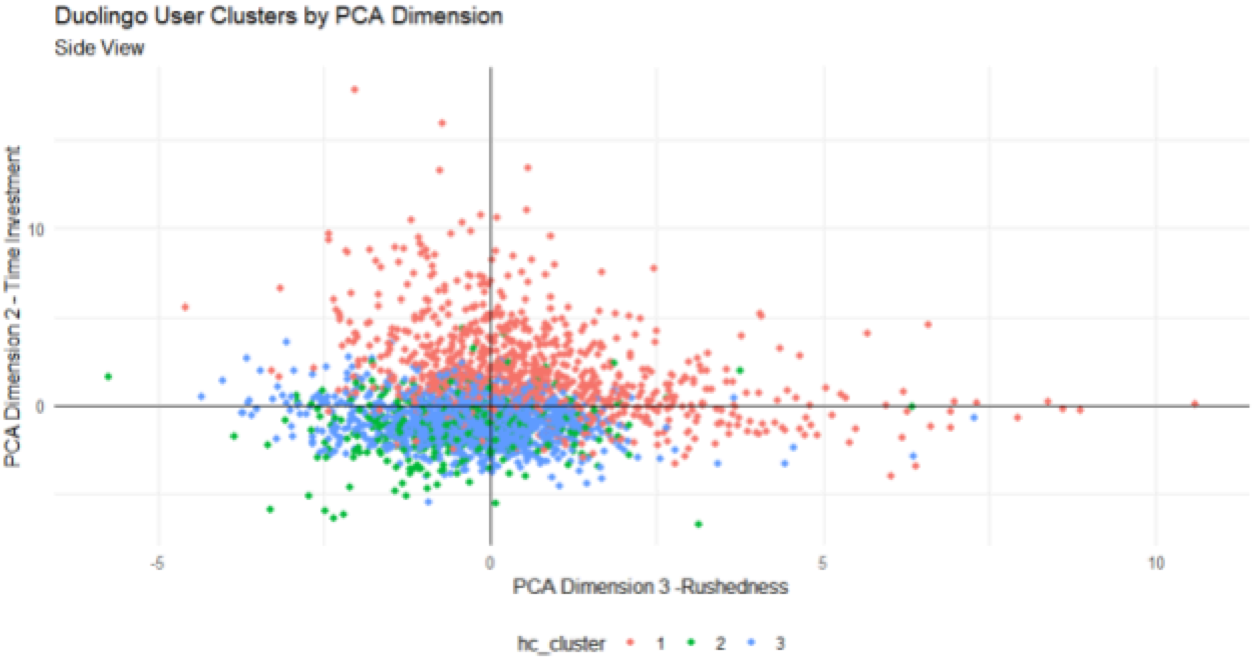


Figure A3 - Cluster loadings by variable

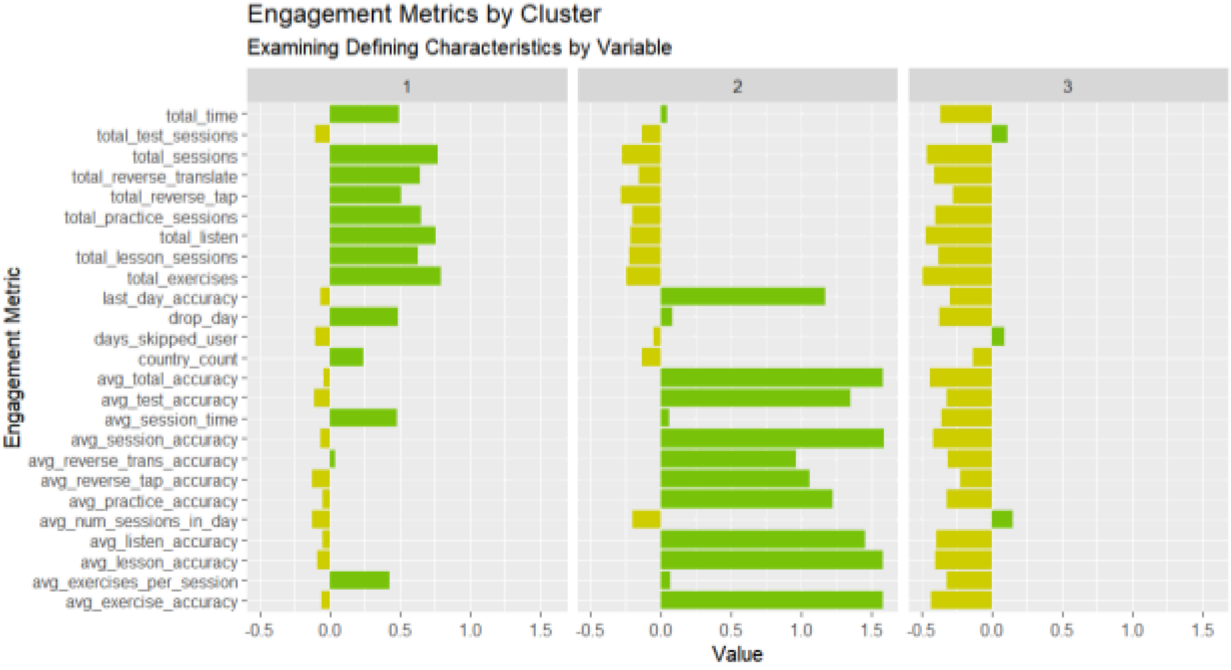
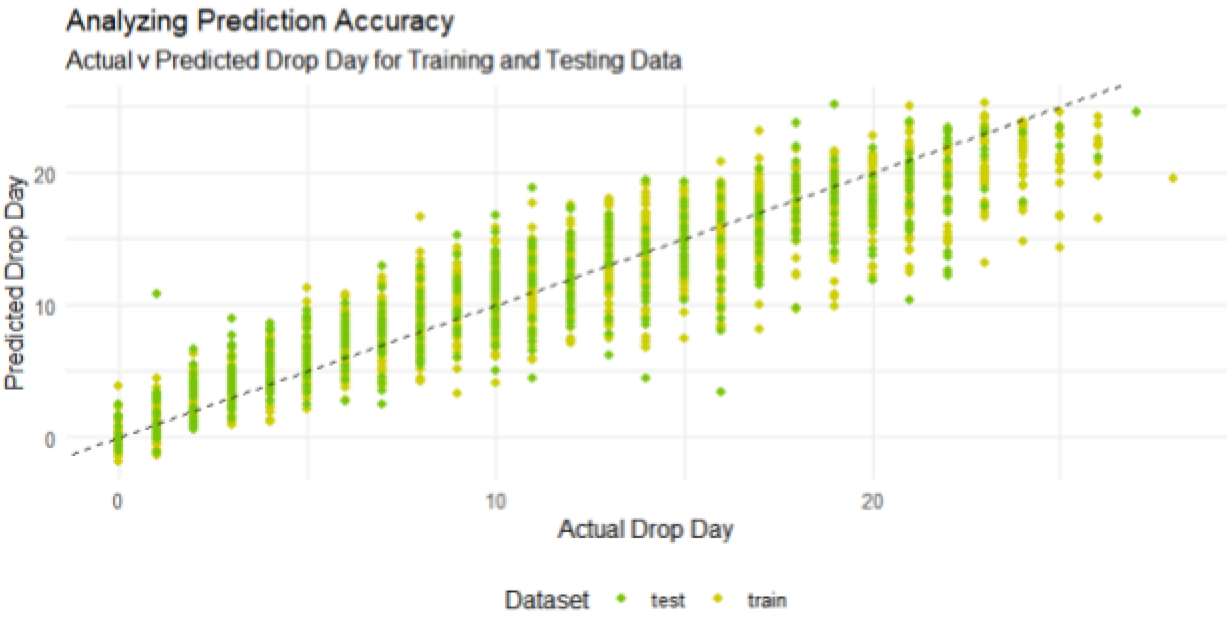


Figure A4 - Scatterplot showing the comparison of each user's actual last day versus their last day as predicted by the best performing XGBoost model


Analyzing Prediction Accuracy
Actual v Predicted Drop Day for Training and Testing Data
Predicted Drop Day
Actual Drop Day
Dataset
● test
● train