Duolingo data challenge

Shuoqi Sun

September 2024

1 Methodology

All the tasks explained in this section were done in Python. I start by checking duplicates records in the two datasets, then merging two datasets by matched user IDs. To prepare the dataset for analysis, I then undertook a series of data cleaning steps:

First,I performed outlier detection on numerical features, and either capped extreme outliers(for example, I capped "n active days" by 97, which is the duration of the sample period), or removed the rows(for example, I removed the rows which contain a negative "highest course progress"), depending on their influence on the analysis. Second, I handled missing values by identifying and dropping columns with more than one third of missing data, namely, "daily goals", "motivation followup". Still, approximately 29% of the rows have missing values more or less, so instead of removing the rows with missing values, I perform imputation to address the issue. To reduce the bias resulted from simple imputation and to preserve contextual information as much as possible, I applied K-Nearest Neighbour imputation for both numerical and categorical features.

For the categorical variables, I applied one-hot encoding for nominal variables(namely, gender, country, platform used, subscription status among others) and ordinal encoding for ordinal variables(namely, income, age, commitment level, proficiency level, among others) to convert them into numerical format. Additionally, I standardized numerical features using MinMaxScaler from Scikit-learn to ensure that all features were on a comparable scale, which is particularly important for KNN method.

I then used the KNNImputer from scikit-learn, specifying an appropriate number of neighbors (k=5) to estimate the missing values based on the average of the nearest data points in the feature space. After completing the imputation, I reviewed the dataset to ensure that all missing values were properly filled and validated the results by checking that the distribution of the imputed values aligned well with the overall dataset.

As the next step in the data-prepossessing pipeline, I engineered several features to make the input data more suitable for this clustering task. First, I constructed a feature called "completion rate" by taking the ratio of number of lessons completed to the number of lessons started, which is a reasonable measure of the user's level of follow-through. For ease of interpretation, I compressed the various responses of "primary language motivation" into two groups: "extrinsic" and "intrinsic", which serves as a simple classification of language learners' motivation. Additionally, I transformed skewed variables, namely "longest streak" and "highest course progress", using log transformations to reduce the impact of extreme values on the model. Finally, I excluded features that are not directly relevant for my clustering task, such as "time spend seconds","future contact".

The resulting dataset is ready to be fed into K-means clustering algorithm, which I chose for its effectiveness in performing unsupervised learning tasks. I used Elbow method to determine the optimal number of clusters. Both 3 and 4 seem to be feasible choices for K, per the elbow graph shows. I decided to go with 3 since it has a higher sihouette score. I then fit the K-means model to the processed dataset, which assigned each data point to its nearest cluster based on Euclidean distance. After clustering, I visualized the clusters along various dimensions to understand the key characteristics of each segment. See below for a graph.

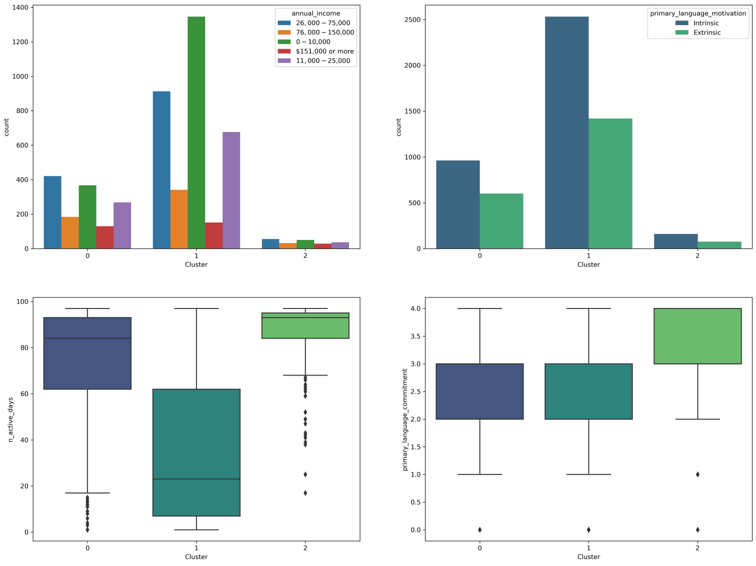

Figure 1: Comparison of salient features across clusters.

2 Description of personas

Based on the outcomes from k-means, below I give an description of the persona from each cluster.

Cluster 0: Goal-oriented Grace(20%)

Grace is a well-paid professional in her late 30s. She uses Duolingo to gain proficiency in a new language, or improves her proficiency on a language she knows, for career advancement. She logs in daily, prefers structured lessons, and is likely to upgrade to premium for exclusive content.

Cluster 1: Young Explorer Daniel(70%)

Daniel is a full-time college student majoring in international relations. He has a genuine interest in learning new languages,and he is always eager to try various language learning opportunities. However, his financial resources are limited, and as a typical student, he's juggling multiple responsibilities, including part-time work, classes, and social activities. While Daniel is motivated to learn, he often gets distracted by social media, video games, and other apps.

Cluster 2: Casual Helen(10%)

Helen is a recently retired high school teacher who is looking for ways to keep her mind engaged during her free time. She's always been curious about languages and sees learning a new language as a way to stay sharp and mentally fit. She enjoys leisurely activities like reading, solving puzzles, and traveling. Helen has the time to learn a new language, and she sees herself very committed to learning the language, but she isn't looking for a high-pressure environment. She's willing to invest in premium features that offer a richer, more structured learning experience.

3 Recommendations

- For dedicated users like Grace, offer paid premium content and tools that help Grace achieve her goals faster, such as 1-on-1 tutoring, personalized feedback, and advanced learning materials. Since she is goal-oriented, a goal-setting feature within the app where Grace can define milestones (e.g., pass a language proficiency test, prepare for a business meeting in a foreign language) can also be very helpful for continuous motivation.

- For users like Daniel who are least likely to pay for Duolingo plus, collaborate with schools and universities to offer discounted access to the app as part of their language training programs. Additionally, since Noah is also less likely to stay active on the app, it can be useful to offer personalized notifications highlighting a fun lesson he hasn't yet tried, or sending him limited-time offer for trying Duolingo Plus for free.
- For casual learners like Helen, they may want a premium subscription that offers more flexibility. For example, allow Helen to customize her lessons and set the pace herself. Offering flexibility around time and lesson length will help her feel in control of her learning without the pressure of a daily schedule.

  Thanks for reading:)