User Segmentation Analysis

David Lurie

2024-09-27

Background

A survey was conducted among Duolingo users from May 1st to August 5th in 2022, providing us with demographic and qualitative data relating to Duolingo and language learning. Along with more quantiative usage data we can attempt to segment users into personas and groups to aid future marketing and product development work.

The data consist of roughly 6000 survey respondents from 10 different countries. The average respondent was in their mid to late 20s and had been on Duolingo for slightly more than a year at the time of the survey. 18% of respondents were students, while 61% were employed to some degree.

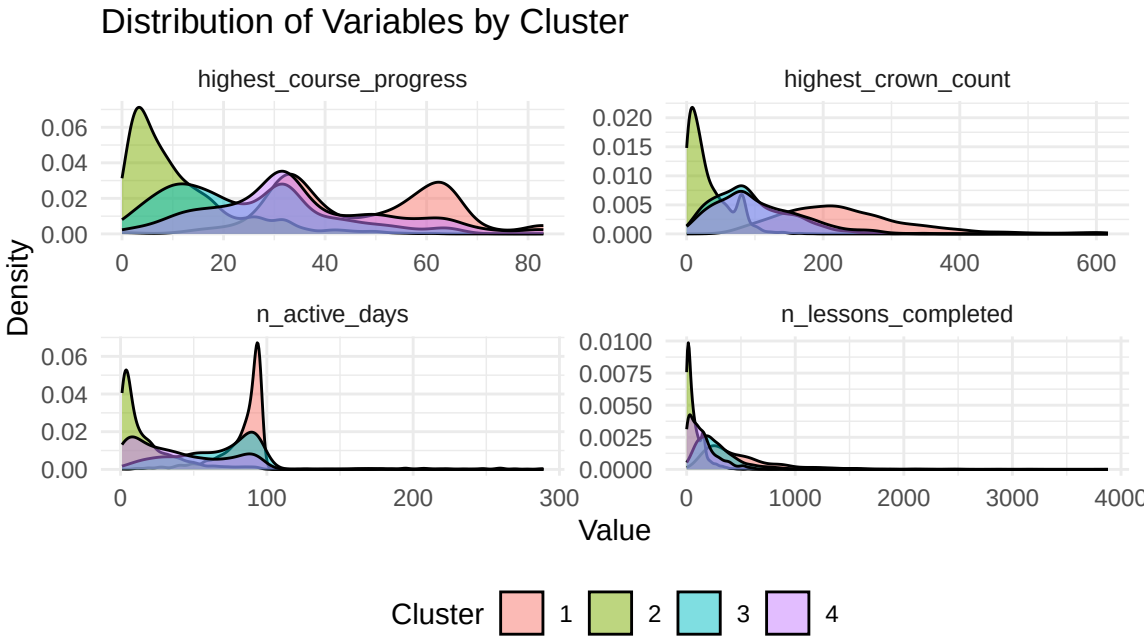
Segmentation

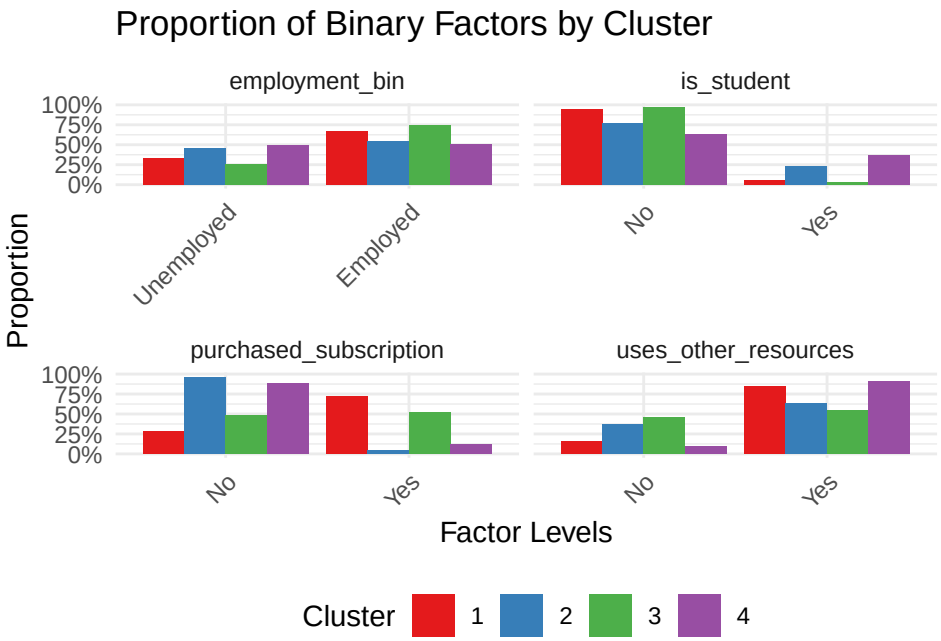
In order to segment the users as precisely as possible, we will use both survey and usage data, and change the encoding of some features so they can be more easily digested by quantitative methods.

Given the high number of features, not all are likely to be relevant. Using a dimension reduction method addresses this while also allowing us to make choices about how many clusters to have and to see how different they are from each other. We will use principal component analysis for its interpretability and relative simplicity.

Of the first two components only about 30% of the total variance in the data is retained, so if we were to use all of the original data instead we might produce very different clusters using the same methods.

To cluster the data we first use hierarchical clustering to find cluster centers, and then K-means in order to ensure stable clusters. The number of clusters was chosen visually with a scree plot.





By plotting the distribution of the four most represented variables in the top two principal components, as well as the clusters against four binary variables, we can gain insights into the segments we have created.

Cluster 1: "motivated learners"

Cluster 1 users are the most likely to have a subscription, and have also been active longer while having completed more lessons. This is a key source of revenue for Duolingo.

Cluster 2: "new learners"

Cluster 2 users are new to Duolingo, and as a result haven't completed many lessons or gained many crowns. They are very unlikely to have purchased a subscription.

Cluster 3: "slower learners"

Like cluster 2, these users are fairly likely to have a subscription but haven't completed as many lessons or gained as many crowns. Many are still newer than cluster 1 users, and could move into that group later. As a high proportion are employed, it's possible they have less time to learn.

Cluster 4: "distracted learners"

Cluster 4 users are most likely to be students and to use other resources. They are very unlikely to have a subscription, but have moved fairly far in a course.

Overall, clusters 1 and 3 are most profitable to Duolingo as they are much more likely than 2 or 4 to buy subscriptions. Cluster 2 contains mostly new users; ensuring that a significant portion of them are retained could eventually see them move into cluster 1 or 3 (i.e. buy a subscription). Cluster 4 might offer opportunities in that they have been on Duolingo for longer while also being unlikely to have a subscription. Trying to understand why they are so likely to use other resources could offer insights into product areas that Duolingo could improve in. In addition, roughly 2 in 5 are students, so targeted efforts like student discounts or free trials for students might see this group increase its conversion rates.