Duolingo Personas

June 1, 2021

Eamon Glackin - Director of Marketing Analytics application

In this report, I will briefly summarize my analysis and present three duolingo *"personas"* which typify three distinct user segments observed from the data. In order to identify these user segments, the survey data and usage data were first cleaned then joined together, and lastly transformed into features usable for unsupervised machine learning techniques such as clustering.

Approach

In order to identify user segments, one key thing to consider is including both survey data as well as usage data. Survey data is helpful for distinguishing between groups based upon demographics, psychographics, and self-reported motivations and asseessments of activity on Duolingo. However, survey responses may not always be a true reflection of the way a particular user or set of users engages with the product. To control for that, the survey data is combined with observational app usage data to better understand how users actually use the product.

After cleaning and joining the data sets, an unsupervised machine learning technique was used to identify distinct clusters. Since there is no particular variable that we are trying to predict and we have no *a priori* beliefs about how many distinct groups of users there ought to be for us to classify users into, clustering techniques are appropriate for this purpose. Typically an algorithm such as ***k-means*** would be used to identify clusters of distinct data points. However, that algorithm is primarily useful for numeric data and has shortcomings when trying to leverage it on categorical data such as survey responses. ***k-modes*** is a similar technique useful for clustering on categorical data, but that would leave us using only the survey data and discarding the numeric app-usage data. Thus, a technique known as ***k-prototypes*** - which is essentially a hybrid of *k-means* and *k-modes*, allowing us to cluster on mixed (i.e. both categorical and numeric) data - was used.

The ***k-prototypes*** algorithm was run several times, with a different number of clusters k for each iteration. After plotting k against a "goodness-of-fit" metric, $k = 3$ was chosen as a reasonable number of user clusters to represent the data (using the "elbow" method to weigh the tradeoff between accuracy and complexity. See figure 1 in the appendix).

Clusters and Personas

The output from the ***k-prototypes*** algorithm with 3 clusters results in the following:

1. **Cluster A** - Typified by the persona *"Boris"*. Users in this segment are typically male, middle-aged, lower-income, and highly engaged with Duolingo but not paying subscribers.

They are intermediate proficiency in the language, and are using the app to review. They need to know the language where they live.

2. **Cluster B** - Typified by the persona *"Yukie"*. Users in this segment are typically older and middle-to-upper income, and may be male or female. They are using the app to learn a language for the first time, not because they need to, but because it is a passion or pursuit - to connect with their culture or heritage (or to prepare for a trip). Users in this group are typically iPhone users and pay for a subscription.
3. **Cluster C** - Typified by the persona *"Olivia"*. Users in this segment are typically younger, female, and lower-income. They are likely to be from an english-speaking country, and are using the app to learn a language for the first time. However, unlike users in cluster A, these users are less committed to learning the language, and engage with Duolingo less frequently.

See `ClusterSummary.csv` for a summary of each cluster, or the notebook 4_PostClustering_DescriptiveStats for charts demonstrating the cluster breakdowns across various metrics such as by age, income, and active rates. For two examples, see figures 2 and 3 in the appendix.

Cluster A - *Boris* Cluster A is typified by the user persona *"Boris"*. He is a user from Russia who is likely in one of the bottom 2 income brackets. He is likely male, middle-aged, and uses an android phone. His survey responses indicate that he is very committed to learning the language and that he uses the app Daily. His actual usage statistics during the observation window largely supports this, as he was active on the platform 2 out of every 3 days from August to November 2018, on average. He typically takes 2.9 lessons per day, has been on the platform for about 2 years, and has made good progress - having completed about 35 rows on average. He has studied the language before and is using Duolingo to review so he can speak the local language where he lives. He does not pay for a subscription, though.

Marketing campaign or product hook - *"Use it or lose it! Need a language refresher? Duolingo is here to help. The best way to stay fluent in a language is to practice using it consistently. With just a few minutes a day, Duolingo will help you learn new words and expressions, and retain what you already know."* Key things to highlight are the convenience and ease of use of having a language app right on whatever device you have, the deep catalog of lessons for all proficiency levels, and the features that promote consistency such as streaks.

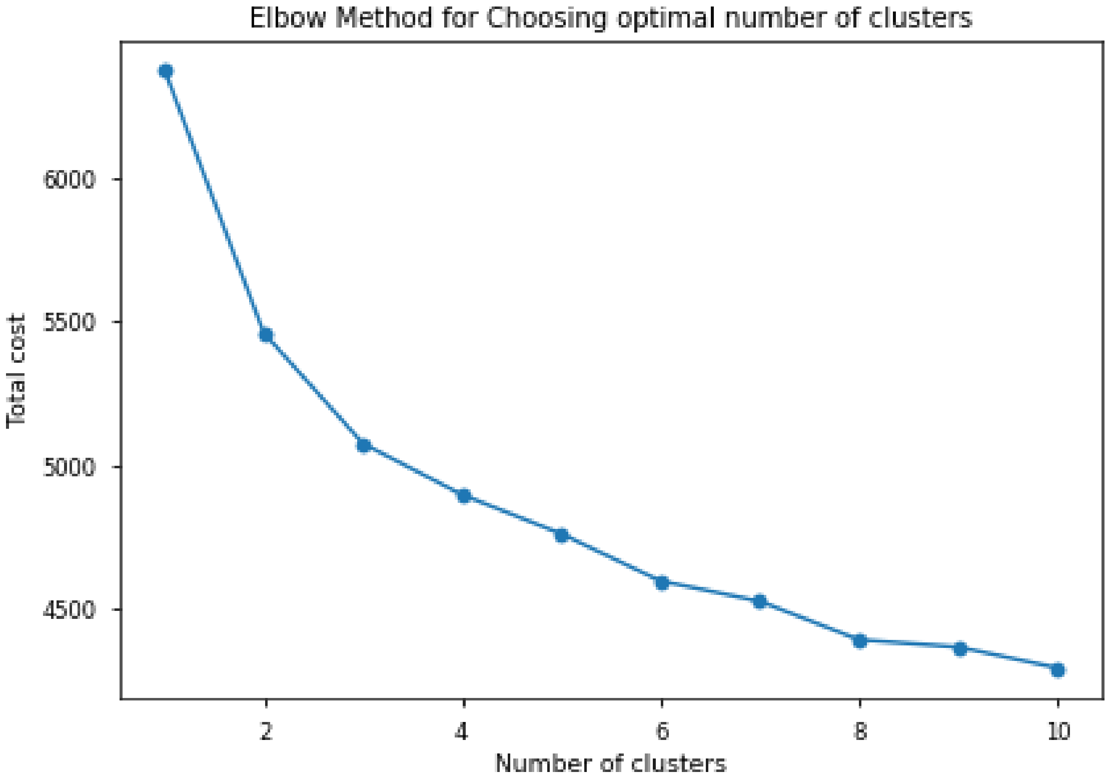
Cluster B - *Yukie* Cluster B is typified by the user persona *"Yukie"*. She is likely an older, wealthier woman from a country like Japan who is learning a language for the first time in order to connect to her heritage or identity. (Or she may be learning a language to prepare for a leisure trip). She is a beginner, learning the language for the first time. She does not need to learn this language for work or to learn the language where she lives, but she is nonetheless very committed to learning the language. This is a passion, or a leisure activity she is dedicated to. She is active more than 75% of days, on average, taking more than 3 lessons per day. She has the financial means to both use Duolingo primarily on an iOS device and to pay for a subscription.

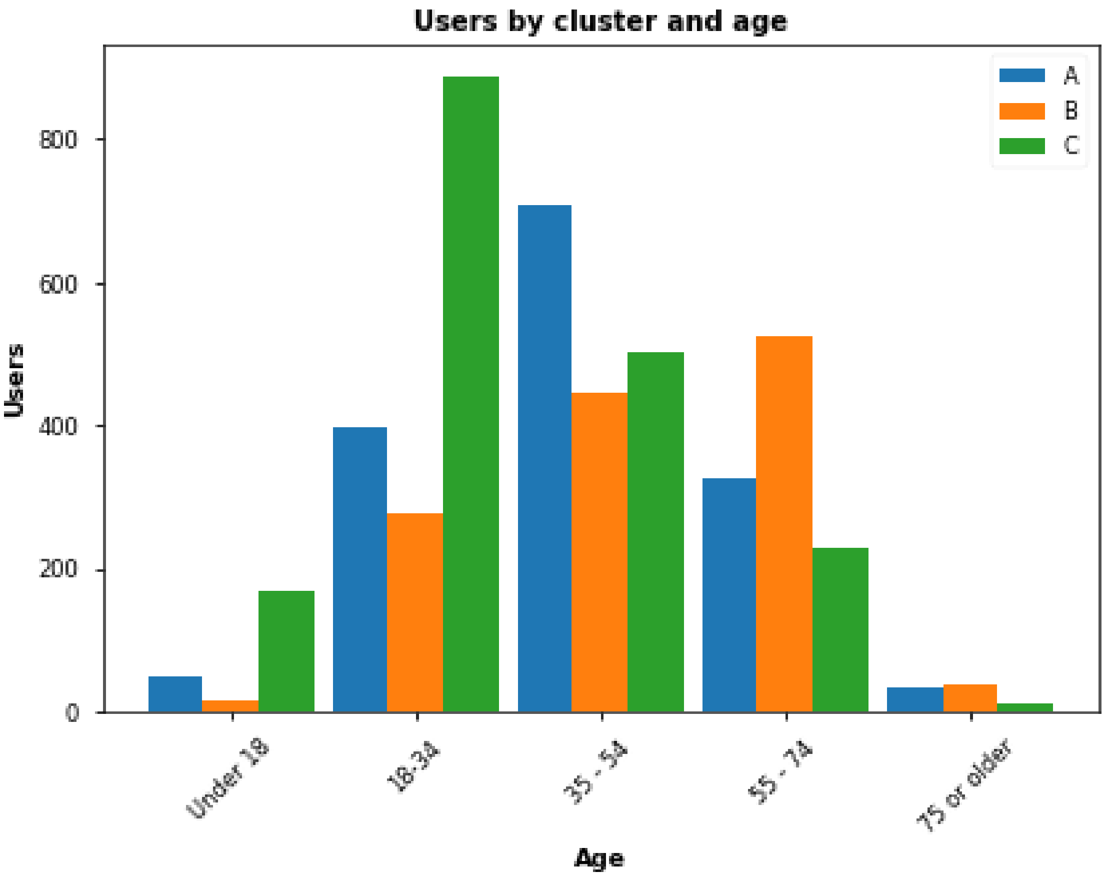
Marketing campaign or product hook - *"Learning a language is more than just learning to read, write, and speak. It is a way to immerse onself in a culture, and how it expresses itself through language. Language is more than communication, it is the means through with culture and heritage are passed down from one generation to the next."* For Yukie, Duolingo is not a means to a professional end nor an economic imperative. She doesn't need to learn a language for work or for daily life. Rather, she is learning a language as a leisure activity or passion project, to connect to her heritage or prepare for a trip. Marketing to Yukie should highlight cultural lessons such as idioms, art, and history. Since she is higher-income and likely willing to pay for a subscription, the paid product should be put front-and-center in any marketing to her. Things to highlight include achievements, streaks, custom app icons, and information about how you are likelier to successfully learn a language by making a financial commitment to doing so.

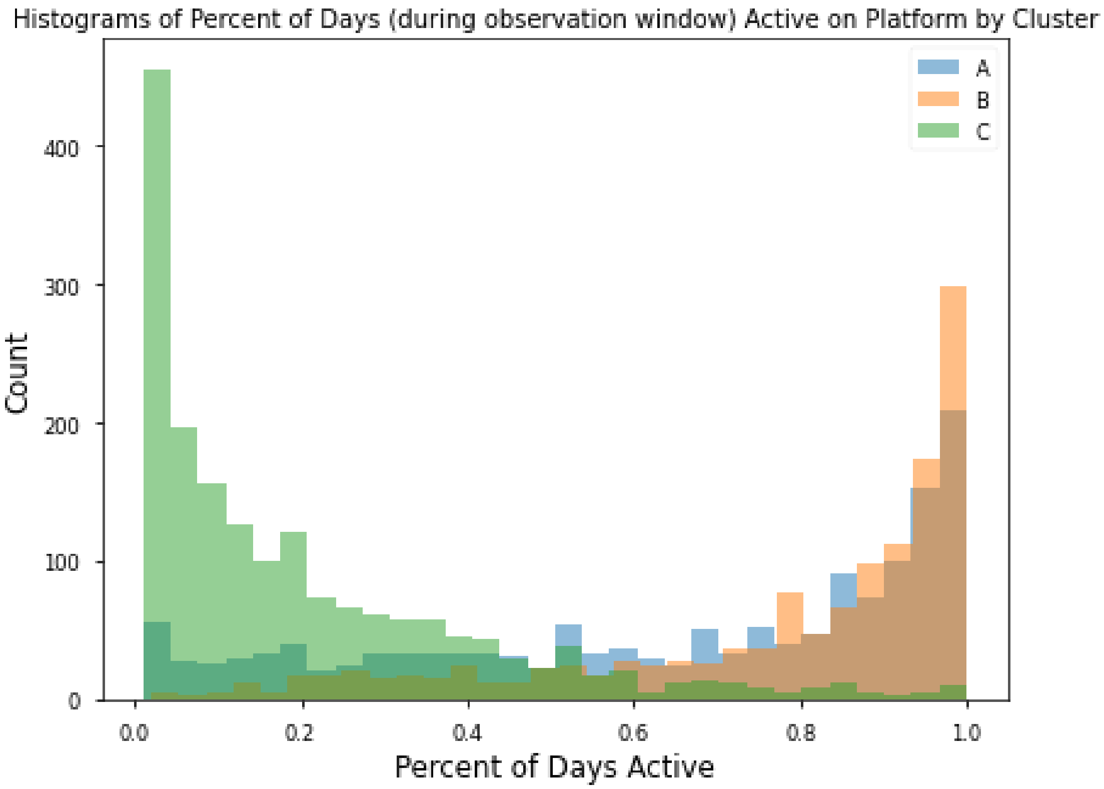
Cluster C - *Olivia* Cluster C is typified by the user persona *"Olivia"*. She is from a similar economic background as Boris from cluster A - very likely in the bottom 2 income brackets and likely to use an android device - but is much more likely to be in an english-speaking country like the UK or US. She is younger, and has only been on the platform for 1-1.5 years (or is likely brand new). Unlike Boris, she is using the app to learn a language for the first time, but she is only moderately committed to learning the language. In her survey responses she is likely to say she uses the app daily, but the usage data suggests otherwise. On average, she only uses the app every 1 in 5 days, or slightly more than once per week. When she does use the app, she takes a similar number of lessons at 2.9 per day. Like Boris, she does not pay for a subscription.

Marketing campaign or product hook - *"Its easier than you think to learn a new language! With just 5 minutes a day, Duolingo can teach you to speak like a local before you know it."* For Olivia, commitment and consistency seem difficult. She is younger and likely distracted by other things in life such as building a career or starting a family. She has aspirations to use Duolingo daily to learn a language, but has trouble finding time to do so consistently. Marketing campaigns, push notifications (if she is already a user), and other communications need to emphasize that lessons on Duolingo are bite-sized and that you can make progress with a low time commitment.

Appendix


Elbow Method for Choosing optimal number of clusters
Total cost
6000
5500
5000
4500
2
4
6
8
10
Number of clusters


Users by cluster and age
A
B
C
800
600
400
200
0
Users
Under 18
18-34
35 - 54
55 - 74
75 or older
Age


Histograms of Percent of Days (during observation window) Active on Platform by Cluster
A
B
C
Count
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
Percent of Days Active