**Duolingo user segmentation**

1. **Summary**

This project aims to develop user segments to aid future marketing strategies and to improve products. We combine the user survey and the usage data to understand the underlying user patterns. Using an unsupervised clustering model, we successfully divide users into two groups. We explore the behavior differences and compare the statistics of each group. We present several ideas that can be relevant for making informed decisions.

**2. Datasets**

**a) Survey data**

The survey dataset, prepared in a csv format, was read in as a DataFrame called survey*.* The 18 attributes of survey include

Classical demographic categories for user segmentation:

* Age, income, country, gender, student, employment status

Quantitative categories measuring their interaction with Duolingo products:

* Platform, subscription status, future contact permission, usage frequency

Qualitative categories measuring their interaction with Duolingo products:

* Commitment level, previous knowledge of language, other resources, motivation for learning

Metadata of the survey:

* Completion of survey, time spent on survey.

As survey responses were entered by users, there are inherently many missing values. We also note that some continuous features (i.e., age, income) were already binned in groups. Each observation comes with *almost unique* user-id’s. Among 6187 observations, there are 37 duplicate user-id’s. Each pair of duplicated users submitted vastly different survey responses. We suspect multiple users may be sharing the same account. Some questions in the survey are free-response types, and these answers required Natural Language Processing to gain useful insights.

**b) Usage data**

The next dataset usage is also provided in a csv format, and contains the user usage information. The user-ids in usage are in the same format as in survey. The user-id’s are later used to join two datasets. Most of the features are straightforward. Here are rephrased features.

* Start date, the number of active days, pre-set daily goal, subscription status, and various achievements in the app.

This data seems to be automatically collected from the system. It contains few missing or invalid values.

**3. Data Cleaning**

Here is an overview of the different approaches to handling missing values and outliers. The methods vary depending on the features. In each case, we provide appropriate assumptions.

* Averages: For features like age or income, the median values of the distributions are used.
* Common sense: If there is no recorded response for usage frequency, we assume the user seldom interacts with the app.
* Maintaining distributions: Non-missing values of the platform [Android, iOS, Web] had the ratio [44:42:14]. To fill each missing platform value, we randomly choose a platform from this [44:42:14] distribution.
* Capping: The middle quartile of the time spent on survey has the approximate range 4-8 minutes. However, negative and over 24-hour completion time were observed. To handle the outliers, we cap the values to range from 0 to 60 minutes. These correspond to 2-97% quantile range. We use the generous 60 minutes cap as we do not want to excessively exclude users. There can be users who take long time providing lengthy and valuable comments, or users who need extra time with typing.
* Invalid values: Some features, such as the course progress, should take non-negative values. These were set to 0.

We process text data from free-response columns to convert each observation as a list of processed, important words. For example,

Original: It's been a lifelong goal to learn this language,I want to learn as many languages as I can

Post-processing: ['lifelong', 'goal', 'learn', 'language', 'want', 'learn', 'many', 'language']

In particular, we correct whitespace, remove punctuation marks, exclude *stopwords*, convert cases, and extract the base of words. Stopwords are commonly used word such as “the”, “a”, “I’ve”.

**4. Feature Engineering and Modeling**

We convert all categorical variables and lists of words to numerical values. We also transform the variables to be more equally distributed, and to be in the same scale. This is an important preprocessing step before applying KMeans clustering algorithm.

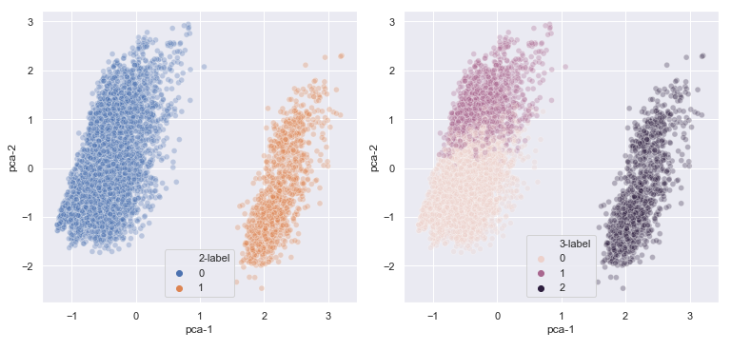
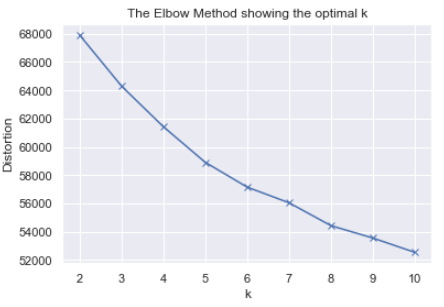
* To transform text data to a meaningful representation of numbers, we use the pre-exiting tool Tf-Idf Vectorizer.
* Nominal categorical variables such as countries are one-hot encoded.
* Ordinal categorical variables:

Some features, such as the commitment level, can be represented in 1-5 scale.

When appropriate, we also use the weighed numbers. For usage frequency, [Daily: Weekly: Monthly] were assigned to [30, 7, 1] to reflect the intrinsic distance between the values.

* Most of the features are not normally or uniformly distributed. For standardization and normalization, we use QuantileTransfomer.

We apply KMeans with the varying number of clusters. Finding the optimal number of clusters is a key part of the process. However, the well-known elbow method did not provide much insight on the number of clusters. We instead use the visualization to find the optimal value. Using the dimensionality reduction tools, tSNE and PCA, we plot the 168-dimension data in the 2-dimenisonal plane. We decide that it is appropriate to segment users into 2 clusters.



**5. Observations**

According to the KMeans model, 5102 users are identified to be in Group 0 and other 977 users are identified to be in Group 1. The average values of motivation level, commitment level, and other factors indicating active engagement are all observed higher in Group 1 (To further back up this conclusion, it is recommended to run statistical tests to compute the likelihood of observed average gaps). The subscription rate is higher, but it is not significantly higher in Group 1. Group 1 can be considered loyal, more profitable, or at least potentially more profitable users.

Some noticeable and interpretable characteristics of Group 1 users are as follows.

1. Many of them have intermediate or advanced proficiency in that language, and took placement test.
2. Students are more likely to be in Group 1.
3. Users from Japan, France, Germany, Mexico, Colombia have a higher likelihood to be in Group 1.
4. Teens and young adults.
5. Male users.
6. They either have very low income or very high income. We wonder if there are explainable correlation between 1), 2), 3), 4) and 6).

Low income users can be students, or users from countries with lower GDP. They may be new immigrants or refugees who have not yet settled down for stable income.

High income users may be the ones who have time and resources to learn many languages or who can afford to travel often.

From text corpus we created, we learn that users watch movies and TV shows to learn languages. Quite many users also reported that they would like to learn new languages for upcoming trips. Some answered that they are immigrants and refugees.

Combining all the observations, we suggest

* Make the placement test easily accessible, fun and engaging.
* Continue to work with schools and support teachers and students.
* If there is a way to know in advance (such as Cache data from websites) travel plans of a potential user, use that to show relevant ads.
* Continue to invest in language modules in Japanese, French, German, Spanish.
* Include phrases from movies and books.