User Segmentation Report

By Wanxin Luo

Methodology

1. **Understand and Explore the Two Datasets**

With the **survey_data** and **survey_users_app_usage** two datasets, I found that there are duplicate records for individual users. To retain only the most recent entries, duplicates in the **survey_users_app_usage** dataset were removed based on the `'start_date'`. Similarly, for the **survey_data**, duplicates were dropped by keeping the last entry.

2. **Exploratory Data Analysis (EDA)**

Before clustering, an exploratory data analysis (EDA) was conducted to visualize the distribution of the entire user base. Later I will compare the distributions of users within each clusters with the distributions of entire user to understand how each cluster differs from the overall population.

3. **Data Cleaning and Preprocessing**

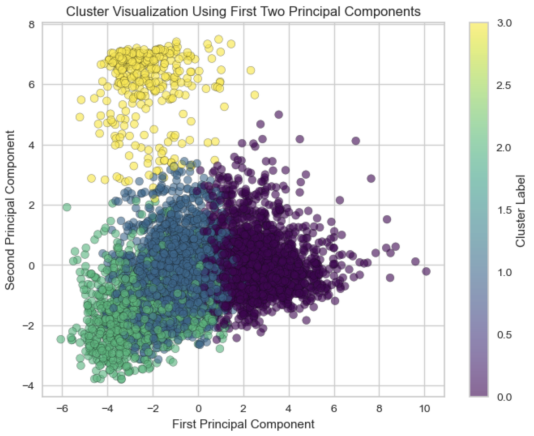
- For categorical features with less than 10% missing values, imputation was performed using the mode of the available data. Then I performed one-hot or ordinal encoding based on their features' atributes.
- For categorical features with more than 10% missing values, missing values were assigned as 'Unknown' without imputing potentially misleading values. Then I performed one-hot encoding to them.
- For numeric variables with missing values, null values were replaced by 0.
- I used a pipeline to perform above imputations and encodings, and I finally standardize all the features to prepare for PCA and K-Means.

4. **Principal Component Analysis (PCA) for Dimensionality Reduction**

After encoding, the preprocessed data contained 74 features, so I performed PCA to simplify the dataset while retaining its core information. The dimensionality was reduced from 74 to 6. The number of the principle component is determined by Scree Plot.

5. **K-Means Clustering**

With the dimensionality reduced, K-Means was chosen as the clustering algorithm due to its efficacy and simplicity in identifying distinct user groups. The optimal number of clusters (`k`) was determined through both the Elbow Method and Silhouette Analysis. These techniques revealed that `k=4` clusters provided the best trade-off between within-cluster cohesion and between-cluster separation. In this case, I segmented users into four distinct personas.



6. **User Segments Statistical Analysis**

I created mean/mode summary tables and distribution visualizations for 4 user groups. Each of the four user segments was characterized by analyzing the distribution of features within the clusters and comparing them to the overall user base. This analysis revealed unique patterns in subscription history, purchasing power, app usage, and language learning motivations across the segments.

- **User Segment 1**: Dedicated, achievement-oriented learners who are actively using this product to achieve their language learning objectives. They are more likely to be financially stable and utilize the premium features of the app.

	Entire Users	Segment 1
Percentage of users who have active subscription during sample period	31.52%	74.80%
Percentage of users who previously or currently subscribed PLUS	30.22%	73.00%
Percentage of users whose annual income > 10000	64.77%	85.40%
Percentage of users who are very or extremely committed to learn this language	53.66%	75.06%
Percentage of users with highest crown count > 150 during sample period	40.45%	72.08%
Percentage of users with active days > 50 during sample period	47.11%	83.51%
Percentage of users with completed lessons > 150 during sample period	49.19%	83.37%

- **User Segment 2**: Users who are less engaged with the product platform. They might be more casual learners or those who are exploring language learning without a firm goal in mind.

	Entire Users	Segment 2
Percentage of users who have active subscription during sample period	31.52%	5.72%
Percentage of users who previously or currently subscribed PLUS	30.22%	3.65%
Percentage of users who are very or extremely committed to learn this language	53.66%	39.45%
Percentage of users with active days > 50 during sample period	47.11%	29.59%
Percentage of users with completed lessons > 150 during sample period	49.19%	32.59%

- **User Segment 3**: Younger, budget-conscious students with moderate engagement and commitment.

	Entire Users	Segment 3

Percentage of users who have active subscription during sample period	31.52%	11.85%
Percentage of users who previously or currently subscribed PLUS	30.22%	8.85%
Percentage of users who are students	27.12%	82.84%
Percentage of users whose annual income < 10000	35.23%	88.49%
Percentage of users who are very or extremely committed to learn this language	53.66%	46.23%
Percentage of users with active days > 50 during sample period	47.11%	21.64%
Percentage of users with completed lessons > 150 during sample period	49.19%	26.15%

- **User Segment 4:** users who didn't complete the survey

	Entire Users	Segment 4
Percentage of users who complete the survey	92.56%	9.74%

Product recommendations

For user segment 1, we can

- Develop advanced features like personalized learning paths, grammar clinics, and proficiency tests to cater to their goal-oriented learning style.
- Introduce mentorship programs to leverage their commitment to language learning.
- Provide loyalty discounts or rewards for consistent app usage and course completion for user retention.

For user segment 2, we can

- Offer short, engaging lessons for users who prefer casual learning.
- Increase gamified elements to make daily practice more appealing.
- Market the effectiveness of free features while promoting the benefits of PLUS.

For user segment 3, we can

- Offer special pricing or a student version of PLUS with features tailored for academic use.
- Collaborate with educational institutions to integrate the product into their language curriculum.
- Align promotions with the academic calendar, such as back-to-school specials.

For user segment 4, we can

- Develop push notifications or email reminders to engage with the app.
- Offer incentives for completing surveys to increase engagement.