A Report for Feature Engineering and User Segmentation

This report describes how I do features engineering from a transactions data and create user segmentation from the features.

Big Five Personality

Big Five Personality includes openness, conscientiousness, extraversion, agreeableness, and neuroticism. To get the features for these 5 traits, I generate a score for each trait. The score will be from 0 (low) to 1 (high). For example, if a person has score 0.9 in openness, it means that he/she has strong desires to explore the world and welcomes new experiences and opinions.

The following subsections explains how the scores will be generated for each trait.

Openness

Openness score starts from 0.5 (neutral). It will be influenced by:

- Number of merchants each user have purchased from. Higher number of merchants means that a user is more open to try out different stores. As a result, a user with higher number of merchants will get higher openness score.
- Payment method. A user uses ‘balance’ method will have slightly higher openness score, compared to users using ‘credit card’. I assume ‘balance’ is a money in a digital wallet. A user that uses digital wallet is considered to be new adopter because it’s a new technology.
- Loyalty program. Users that joins loyalty program are considered to be less open. As a consequence, they will get less openness score.

Conscientiousness

Individuals with high score in conscientiousness are well-organized people and prefer structure in their lives. They are reliable and consistent, take their finances seriously, and hardly ever go over budget.

Conscientiousness score starts from 0.5 (neutral). It will be influenced by:

- Stability of total transaction amount per week. A user whose expenses each week more or less stays constant will get higher score.
- Transaction note. Users that puts a note in transaction shows that they are more organized. They want to know the detail of things and keep the information for later usage. Consequently, users that puts notes will get higher score.

Extroversion

Individuals with high score in extroversion love interacting with others. Extroverts are more susceptible to impulsive purchases because the people they interact with e.g. friends, store workers can influence them to do unplanned big purchases.

Extroversion score starts from 0.5 (neutral). It will be influenced by:

- Number of impulsive purchases. To detect impulsive purchase, I collect how many purchases with total amount is higher than the sum of mean and standard deviation of all transactions. A user with higher number of impulsive purchases will get higher extroversion score.

Agreeableness

Agreeable people seek products that align with their values and are approved by their peers when making purchase decisions.

Agreeableness score starts from 0.5 (neutral). It will be influenced by:

- The popularity of the merchant. One possible reason user buys product from popular merchants is they want to be approved by their peers. As a result, user that always purchase from popular merchants will get higher agreeableness score. On the other hand, user that purchases from not popular merchants will get lower score. To detect if a merchant is popular, I calculate how many users a merchant has. Greater number of users a merchant has, it's considered to be more popular. I group the popularity into 4 types: very popular, popular, normal, not popular.

Neuroticism

Individuals who score high on this trait have emotional instability and frequent mood swings.

Neuroticism score starts from 0.5 (neutral). It will be influenced by:

- Number of refunded transactions. Users that do refund tend to have emotional instability. They change their minds after buying something. Consequently, if a user ever does refund more than once, he/she will get slightly higher neuroticism score.
- Number of transactions with merchant rating one (very bad). Users that give rating one are prone to have emotional instability. They are more sensitive and reactive to things. As a result, users that always give rating one will get higher score in neuroticism.

Work and Home Location

I utilize the transaction location (latitude and longitude) to define work and home locations.

Assumptions:

- If users do online transaction, they are more likely to do it from home or office.
- If they do in-store transaction, their office or home are more likely located near to the store.

Steps:

1. Firstly, I filter the transactions for each user.
2. Then, I cluster the transaction locations using DBSCAN algorithm. I order the clusters based on the number of transactions. The cluster with highest number of transactions will be the first order. I take the first order as the cluster for home location and second order as the cluster for work location. If DBSCAN only output one cluster, I consider work and home location are near to each other.
3. To generate the latitude and longitude of work and home location, I calculate the mean of latitude and longitude from transactions in the cluster (work location cluster and home location cluster). It's probably better idea to get the location based on the centroid of the cluster.

Gender

The followings are steps to generate gender feature:

1. Firstly, I categorize if a merchant sells products for male, female, or both gender. I utilize Large Language Model (LLM) to categorize the merchants.
2. For each user, I collect a list of merchants and their product's gender. If a user interact more with merchants that sell male product, the user's gender is considered to be male. Otherwise, the gender is female.
3. After doing previous approach, gender of 1114 users still cannot be defined because they always interact with all-gender merchants. To solve that, I try to utilize the merchant category atttribute.
4. For each merchant category, I calculate the percentage of male-only merchant, female-only merchant, and all-gender merchant. Then, I use this percentage information as a probability to generate the gender information of each merchant category. If a user interact more with merchant categories that sell male product, the user's gender is considered to be male. Otherwise, the gender is female.

User Segmentation

I create a user segmentation based on average expense per user and big five personality. I found a quite interesting segment based on agreeableness score and average expense per user. Three user segments are generated from the data (see figure 1). It shows that users with higher agreeableness score have low average expenses. On the other hand, users with lower agreeableness score have low average expenses and high average expenses. Apparently, the other personality traits cannot create meaningful segments from average expense.

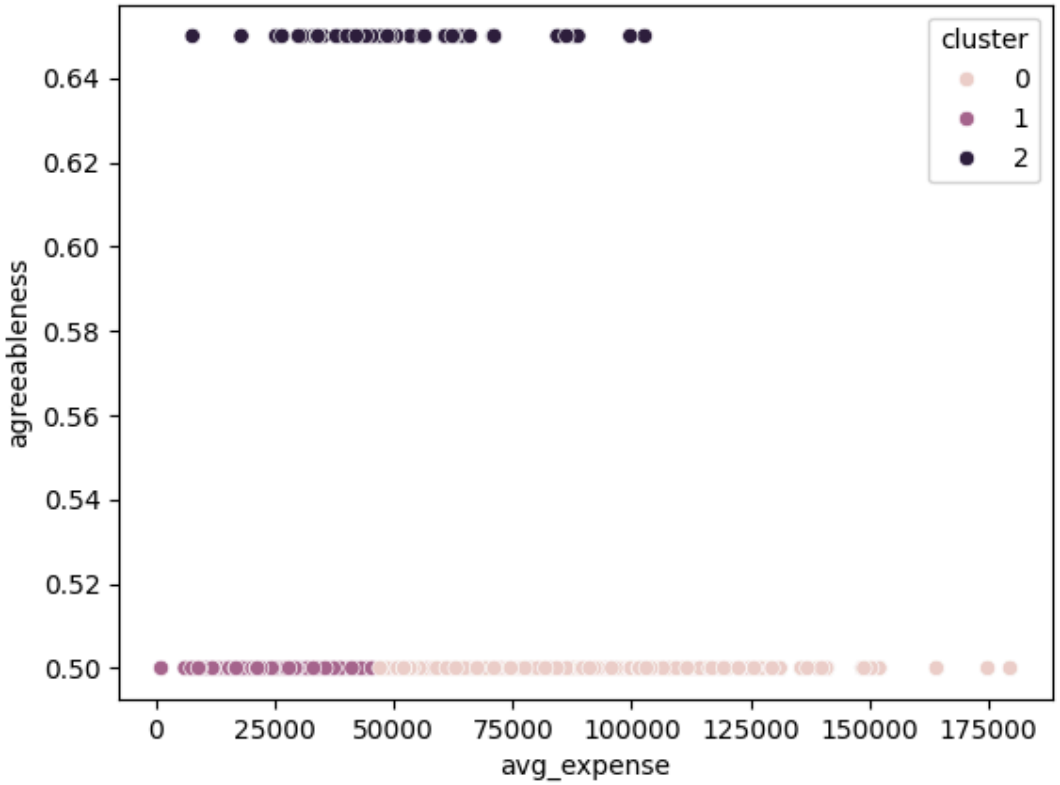

Figure 1. User segmentation based on agreeableness and average expense