STAT 451 Group 2 Proposal

Eva Song, Ani Shi, Jiahua Zhang, Peter Li, Sam Li

Data Description

> Data Source: https://www.kaggle.com/datasets/jmmvutu/ecommerce-users-of-a-french-c2c-fashion-store?select=6M-0K-99K.users.dataset.public.csv

The dataset from kaggle contains user information from a C2C (Consumer to Consumer) fashion e-commerce platform based in France. The platform has over 10 million registered users, making it suitable for analyzing user behavior characteristics on a C2C fashion platform. The data can aid in researching user preferences, consumption habits, activity levels, etc.

The number of rows N = 98913, number of columns p = 24. We will select some of these variables for our project analysis.

Variable	Description	Variable	Description	Variable	Description
`identifierHash`	User ID hash	`type`	Entity type	`Country`	User country (French)
`language`	Preferred language	`socialNbFollowers`	Followers count	`socialNbFollows`	Following count
`socialProductsLiked`	Liked products count	`productsListed`	Current unsold products	`productsSold`	Total sold products
`productsPassRate`	Products quality rate	`productsWished`	Wishlist count	`productsBought`	Purchased products
`gender`	User gender	`civilityGenderId`	Gender as integer	`civilityTitle`	Gender title
`hasAnyApp`	Used any official app	`hasAndroidApp`	Used Android app	`hasIosApp`	Used iOS app
`hasProfilePicture`	Has custom avatar	`daysSinceLastLogin`	Days since last login	`seniority`	Days since registration
`seniorityAsMonths`	Months registered	`seniorityAsYears`	Years registered	`countryCode`	Country (ISO-3166-1)

Research question:

Our goal is to conduct user analysis of the C2C fashion e-commerce platform based on the concept of `user lifecycle`. This includes understanding the characteristics and correlations of users at different stages of life cycle and, further making recommendations for user management and marketing strategies.

User Lifecycle:

- **Acquisition** (Sign Up):
  - *Objective*: Analyze regional marketing strategies by clustering countries based on user activity levels, exploring similarities and differences in user behavior patterns across countries.
  - *Method*: PCA, Clustering (k-Means or DBSCAN)
- **Retention** (Regular Usage):
  - *Objective*: Predict user activity level (usage duration). This information will help in developing personalized marketing strategies and retention efforts.
  - *Method*: Regression models (Linear regression, LASSO, etc.)
- **Revenue** (Monetization):
  - *Objective*: Determine whether a user is more inclined to buy, sell, or has no inclination.
  - *Method*: Classification models (SVM, Decision tree, etc.)

Code to read the data

```
import numpy as np
import pandas as pd
df = pd.read_csv(’https://uwmadison.box.com/shared/static/kver8zowqg58fvyi496gathczucmatt6.csv’)
print(df.shape)
df.head()
```

	identifierHash	type	country	language	socialNbFollowers	socialNbFollows	socialProductsLiked	productsListed	productsSold	productsPassRate
0	-1097895247965112460	user	Royaume-Uni	en	147	10	77	26	174	74.0
1	2347567364561867620	user	Monaco	en	167	8	2	19	170	99.0
2	6870940546848049750	user	France	fr	137	13	60	33	163	94.0
3	-4640272621319568052	user	Etats-Unis	en	131	10	14	122	152	92.0
4	-5175830994878542658	user	Etats-Unis	en	167	8	0	25	125	100.0

5 rows × 24 columns