University of Monastir
Higher School of Computer Science and Mathematics
Computer Science Departement
Business Intelligence

Data analytics of C2C fashion store
7 juin 2021

The data set was scraped from a successful online Customer-to-Customer C2C fashion store with over 9M registered users. The store was first launched in Europe around 2009 then expanded worldwide. Only registered users are included in the data set.

The Owner of the data set is : Jeffrey Mvutu Mabilama
Size :10.27 MB
Displaying 27 columns, 98913 rows in table

Identifierhash : Anonymous unique id
Type : The type of entity. This file contains only "user" entities
Country : User's country (written in french). See also the column "Country Code" if you prefer an ISO identifier.
Language : The user's preferred language (the language of their interface when using the site)
Socialnbfollowers : Number of users who follow this user's activity. New accounts are automatically followed by the store's official accounts
Socialnbfollows : Number of user account this user follows. New accounts are automatically assigned to follow the official partners
Socialproductsliked : Number of products this user liked
Productslisted : Number of currently unsold products that this user has uploaded.
Productssold : Number of products this user has sold
Productspassrate : % of products meeting the product description. (Sold products are reviewed by the store's team before being shipped to the buyer.)
Productswished : Number of products this user added to his/her wishlist.
Productsbought : Number of products this user bought
Gender : user's gender
sex
civilitygenderid : civility as integer
civilitytitle : Civility title
hasanyapp : user has ever used any of the store's official app
hasandroidapp : user has ever used the official Android app
hasiosapp : user has ever used the official iOS app
hasprofilepicture : user has a custom profile picture
dayssincelastlogin : Number of days since the last login. All user data were fetched the same day. See also "seniority".
Seniority : Number of days since the user registered
Senioritysasmonths : see seniority. Here, expressed in months
Senioritysasyears : see seniority. Here, expressed in years
Countrycode : user's country (ISO-3166-1)

Your Work : Data exploration, clustering of the customers, which user’s are likely to leave the store.

Data description

Load data in a data.frame

data=read.table("C:/Users/Asus/Desktop/BusinessIntellignce/caseStudyBI/6M-0K-99K.users.dataset.public.csv",sep=",",header=TRUE)

Dimension of data

dim(data)

44850 observations, users

24 variables

Data types and description

str(data)

Variable	Type	Description
identifierHash	**num**	
type	**chr**	**How many type of users ? unique(data$type)**
country	**chr**	**How many country ?**
language	**chr**	**How many langage ?**
socialNbFollowers	**int**	
socialNbFollows	**int**	
socialProductsLiked	**int**	
productsListed	**int**	
productsSold	**int**	
productsPassRate	**num**	
productsBought	**int**	
gender	**int**	**How many gender ?**
civilityGenderId	**chr**	**How many gender ID ?**
civilityTitle	**chr**	**How many civility title**
hasAnyApp	**chr**	**Binary**
hasAndroidApp	**chr**	
hasIosApp	**chr**	
hasProfilePicture	**chr**	
daysSinceLastLogin	**int**	
seniority senio	**int**	
seniorityAsMonths	**num**	
seniorityAsYears	**num**	
countryCode	**chr**	

Missing values

anyNA(data)

no missing data

Data exploration

Univariate exploration
Bivariate exploration

How do you explore categorical data? Contingency tables, barplots, pie charts, mosaique plots
How do you explore numeric data ? summary boxplot, histograms, ….
How to explore numeric variable relatively to categroical variables

Package ggplot
!!! variables having the same meaning
!!! variables having no statistics values like ID
Comment results and graphs

Data clustering and cluster validation
Data classification and validation

Un compte rendu :
Les résultats et leur analyse.
Ne pas mettre le résultat tel qu'il est présenté dans la console de R. Il faut le présenter autrement. Le compte rendu sera propre.
Un fichier script :
commentaire description
commentaire num figure dans votre compte rendu
puis code

!!! Ne pas mettre le code dans le compte rendu
J'accepte monôme, binôme et trinôme
Attention : l'examen de TP va porter sur la même base de données.

Des recommandations pour la réalisation de l'étude de cas

I. Préparation des données

```
data <- read.csv2(
"6M-0K-99K.users.dataset.public.csv", header=TRUE,
        sep=',')
```

- Donner la dimension des donner en utilisant dim.
- Donner les noms des variables en utilisant names.
- Vérifier s'il y a des valeurs manquantes avec la commande suivante :

```
data[is.na(data)==TRUE
```

- visualiser un aperçu de la base en utilisant la commande str
- décrire les données en précisant le nombre de variables, leurs types, leurs signification
- pour chaque variable catégorique, donner le nombre de valeurs distinctes en utilisant la fonction unique. Par exemple :

```
unique(data$type)
[1] "user"
```

- Pour chaque variable, expliquer la signification de toutes les valeurs prises par les variables catégoriques
- Essayer à ce niveau, de faire des déduction pour réduire le nombre de variables : Une seule valeur, un identifiant : par d'importance statistique, deux variables qui représentent la même information mais avec un codage différents. Décider en fonction de votre observation et compréhension des variables.
- Faire une transformation des variables de type chr en des variables de type Factor si elles prennent un nombre très réduits de valeurs.

> Factors are the data objects which are used to categorize the data and store it as levels. They can store both strings and integers. They are useful in the columns which have a limited number of unique values. Like "Male, "Female" and True, False etc.
> They are useful in data analysis for statistical modeling.
> Factors are created using the factor () function by taking a vector as input.

- Utiliser la commande suivante pour cette transformation. Par exemple:

```
factor(data$language)
factor(data$gender)
```

- Remarquer la variable seniorityAsMonths est de type double donc il faut la transformer en double.

```
seniorityAsMonths  : chr  "106.83" "106.83" "106.83" "106.83" ...
data$seniorityAsMonths=
as.double(data$seniorityAsMonths)
```

- faire la même transformation pour les variables qui sont de type chr et dont les valeurs sont de type double

- A ce niveau, vous allez obtenir une nouvelle base de données après toutes les transformations et après avoir réduit des attributs (20 attributs). Exécuter de nouveau str et observer l'état de votre base de données.
- Calculer la corrélation entre les variables numériques. Observer le résultat et décider de la réduction des attributs au cas où vous trouvez des variables fortement corrélées (>0.9).
- Vous allez obtenir les données avec lesquelles vous allez poursuivre l'analyse 18 attributs
- Faire de nouveau str et essayer d'observer le résultat.

II. Exploration des données

- Calculer le résumé statistique des données avec summary pour les variables numériques. Commenter les valeurs.
- Tracer les histogrammes, les boxplot et les scatter plot pour les variables numériques
- Pour les variables de type Factor, dessiner les barplot

III. Machine learning

- Exécuter des algorithmes de clustering pour voir s'il est possible de catégoriser les clients en des catégories.
- Exécuter arbre de décision ou bien KNN pour décider si le client est susceptible de quitter le store ou bien de rester et d'effectuer des transactions.