T9 - Big Data

T-DAT-901

Recommender

KaDo Project

1.1.2

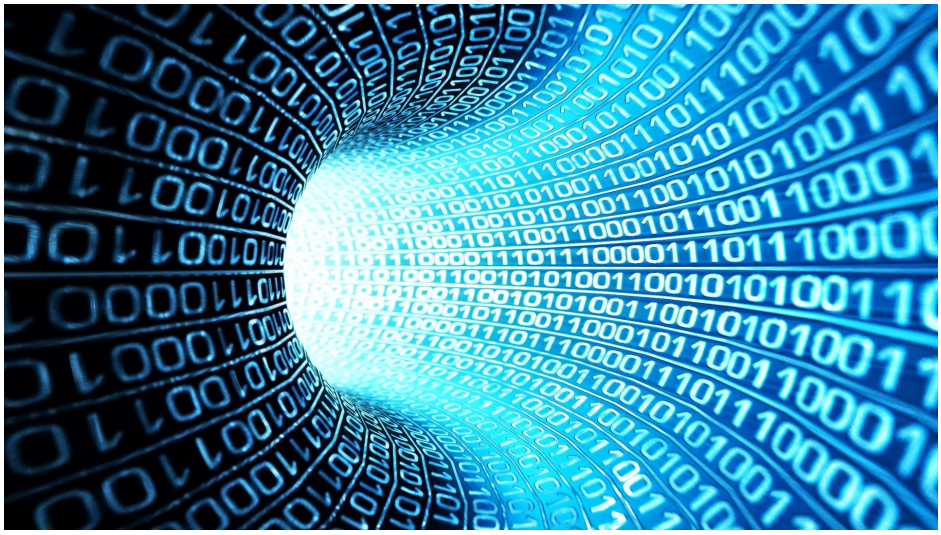
EPITECH.

Recommender

binary name: recommender_*$AcademicYear*_*$GroupNumber*.zip
language: Python or R or anything else that gets the job done
build tool: no need here

> *It's not the customer's job to know what they want.*
>
> *People don't know what they want until you show it to them.*
>
> *Get closer than ever to your customers. So close that you tell them what they need well before they realize it themselves.*
>
> *Steve Jobs*

A company granted you access to KaDo: a database containing millions of bought items, divided in 3 categories. For instance, a bottle of red wine belongs to *Famille*: alcohol, *Univers*: wine, *Maille*: red wine.



In order to increase customers' loyalty, the company expects you to:

- segment the customers to get a clearer insight
- add charts and figures to help this company visualize their customers' profiles
- build a recommender system to offer a gift to each client, based on their preferences
- use different type of recommender system: user based, item based…
- write a kickoff to describe precisely what you will do with the dataset

Tools and techniques

Specificity of the dataset

Most recommender systems are based on grades given by clients on items. This is not your actual situation, as an item is either bought by the customers or not: you have no idea if they like it or not. Despite that, you might want to measure the satisfaction for an item.

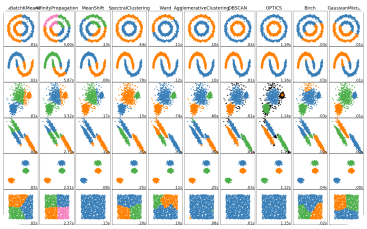
Moreover, keep in mind that you have: hundreds of thousands clients for thousands items, involving that your client/product matrix will be very sparse.

Recommender systems

Many different types exist. Look for intel to get inspiration, or build yours from scratch!

> *Predictive accuracy is substantially improved when blending multiple predictors. Our experience is that most efforts should be concentrated in deriving substantially different approaches, rather than refining a single technique. Consequently, our solution is an ensemble of many methods.*
>
> *2007 Netflix prize winners,*

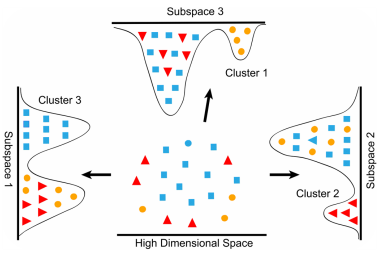
Clustering



A plethora of implementations can be found in datascientist librairies:

- *KMeans*: centroid-based clustering
- *DBSCAN*: density-based clustering
- Hierarchical clustering
- Spectral clustering

Dimension reduction



Approaches can be divided into feature selection and feature extraction:

- **P**rincipal **C**omponents **A**nalysis
- **S**elf **O**rganizing **M**aps

Specifications

Kick-Off

Companies will ask you to do a presentation of your plan, before they give you any data. It's a good way for them to see if they want to hire you.

Based on the information you were given, you're asked to do a presentation in detail of your method and how you intend to solve their problems:

- they want you to build a recommender system to offer one or more items to their clients
- they also need you to cluster their clients to analyze the dynamics of the consumer

You have to use business knowledge and **a priori rules** combining with recommender system algorithmfound below.

Once you complete this step, you will get access to the data, and everything should be easy… Except it will not: you'll encounter unsuspected problems and have to deal with it.

Visualization and data exploration

You're expected to add a descriptive part of the dataset with all relevant statistics and figures, for instance:

- Numbers of items per Maille, Univers, Family
- Most popular items in each category
- Mean price for items in the categories
- Mean and std numbers of items per clients
- Means price spend
- Mean number of items per tickets
- …

All of these statistics will help you to have of ground vision of the topic. It will help you to decide a priori rules which are necessary in that project.

For example, if some *famille*, *maille* or *univers* happened to be far more expensive than other, you might want to recommend those items to spendthrift customers. Knowing whether a category is expensive or not will be based on your ground knowledge

Pre-processing

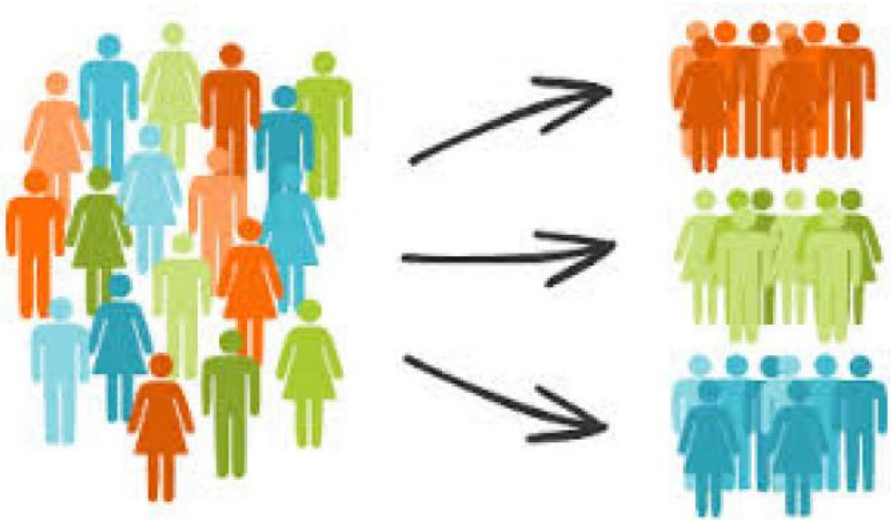
You will have to handle both categorical and numerical variables.

It is your job to detect outliers and put them away for the rest of the analysis. It will need a good knowledge of your dataframe. Also remind that you deal with labelled item, you can look them on the internet.

Segmentation

Rather than just making suggestions, it's important to understand the general dynamics of customers. You are asked to do a client segmentation, which consists in classifying clients in groups depending on their behavior.

This segmentation should be based on concrete client-based attributes.



Deliveries

- a kickoff should be first submitted, in order to get access to the data
- the final delivery should contain:
  - a report to get into specific details of your methods, describing your pre-processing as your recommender system
  - a program `user_recommendation` which take any client id in input then give back:
    - a statictical description of the client
    - a display of the client's segmentation analysis (profil, type of client, general behavior, …)
    - at least one gift for the client **and** an explanation on why this gift was recommended

Bonus

You can improve this project in many ways, including:

- adding the possibility to choose between different recommender systems
- measuring the quality of your recommendation through metrics evaluation
- applying NLP on the items labels
- adapting the gift depending on the date: birthday, Christmas, Saint Valentines, Easter Sunday…
- automating the production of a gift card containing illustrations of the offered product
- exploring the data using VR
- anything else you can think of to sell your project to future clients