1SC4892 – E-Marketing

Instructors: Céline Hudelot, Myriam Tami
Department: DOMINANTE - INFORMATIQUE ET NUMÉRIQUE
Language of instruction: FRANCAIS
Campus: CAMPUS DE PARIS - SACLAY
Workload (HEE): 40
On-site hours (HPE): 27,00

Description

Web datackathon: web data challenge!

E-Marketing - With an industrial partner (2019 : Doctolib - 2020-2021 : Procter & Gamble - 2022 : Rakuten)

The objective is to implement, on concrete data provided by the partner and enrich with web-based data, several of the approaches discussed in the ST with the aim of an application in marketing. It will consist in :

- Analyzing and translating the needs
- Building the data analysis chain: from collection (enrichment) to interpretation and visualization
- Designing the underlying technical architecture
- Evaluating, validating and taking a step back from the product on the solution developed.

Quarter number

ST4

Prerequisites (in terms of CS courses)

Information systems and programming- Algorithms design and complexity - Statistics and machine learning

Syllabus

The precise content of this EI may be subject to change each year, depending on the industrial partner involved. The contents of the three previous editions of this EI are given for information only.

===

EI Doctolib, 2019

Doctolib is a major player in the management of medical appointments, bringing patients and healthcare professionals together. He collects a lot of data, via the web, on doctors and patients, and wonders how best to value them.

Doctolib proposes to focus on dental office attendance and to better predict the presence or absence of patients at scheduled appointments. Indeed, among the patients who book on the Doctolib platform, a certain number do not show up for the appointment. Two cases : either they have cancelled and therefore the doctor is able to use the time slot for another patient or they do not cancel and the time slot is lost. This is the second case that will be the subject of the project. This involves using a data set, provided by Doctolib, to build a model to predict patients who do not attend appointments without first cancelling.

==

EI Procter&Gamble, 2020

OralB is a leading brand in Powered tooth brush category with 20% of people in France using powered tooth brush. Based on the research, powered tooth brush gives a better plaque removal by 21% as compared to manual toothbrush. To be able to attract more consumers to try powered tooth brush, we need a marketing campaign that is able to target the right audience at the right time. Using the data on the consumers that have tried the toothbrush, we could develop the algorithms to find the consumers that have a high likelihood to buy Powered tooth brush. The project aims at leveraging data science to make the BEST business decision regarding precision marketing.

==

EI Procter&Gamble, 2021

Procter & Gamble (P&G) is an American multinational corporation that sells consumer products. P&G offers many well-known brands such as Gilette, Braun, OralB, Ariel, Lenor, Pampers, Always, Head & Shoulders, Herbal Essence, Febrez, Mr Propre, Vicks ... They are currently launching a new product in France called Fairy pods for Auto dishwashers.
In this EI, the goal is to identify, using machine learning algorithms, consumers who are very likely to have automatic dishwashers. This will allow the P&G marketing team to target them in media campaigns and attract as many consumers as possible to try the Fairy Pods, while spending the lowest possible budget. Based on the demographic / behavioral characteristics and web actions of consumers, part of which is provided via a database format, the goal is to predict whether or not they have an automatic dishwasher. Students are thus both encouraged to develop strategies for enriching the

databases provided but also for managing missing data and any other problem that a data scientist is confronted with when working on real data.

==

EI Rakuten, 2022

Rakuten Institute of Technology (RIT) is the incubation center of AI research and technologies that fuel Rakuten’s innovation. It is the core AI research wing of Rakuten, is spread across six geographical locations including Tokyo, Singapore, Boston, San Mateo, Bengaluru, and Paris. RIT does applied research on three major verticals, customer understanding, natural language processing, computer vision. RIT also engages in moonshot initiatives includes programs like curing cancer using AI, quantum computing, and building a fully autonomous mobile network.

Context

This challenge focuses on the topic of large-scale product type code (possibly, multimodal using text and image) classification where the goal is to predict each product’s type code (which defines the category of the product) as defined in the catalog of Rakuten France.

The cataloging of product listings through title and image categorization is a fundamental problem for any e-commerce marketplace, with applications ranging from personalized search and recommendations to query understanding. Manual and rule-based approaches to categorization are not scalable since commercial products are organized in many classes. Deploying multimodal approaches would be a useful technique for e-commerce companies as they have trouble categorizing products given images and labels from merchants and avoid duplication, especially when selling both new and used products from professional and non-professional merchants, like Rakuten does. Advances in this area of research have been limited due to the lack of real data from actual commercial catalogs. The challenge presents several interesting research aspects due to the intrinsic noisy nature of the product labels and images, the size of modern e-commerce catalogs, and the typical unbalanced data distribution.

Task definition

Participants are required to design a classifier to categorize products in the Rakuten France catalog into product type codes.

Class components (lecture, labs, etc.)

The work will be done, in competition or challenge mode. Students will be divided into groups of 4 to 6 people and will work in groups to propose the best prediction model and a product view of the proposed model. The proposed model will be evaluated on a test set that will be provided at the end of the week to the different groups.

Grading

The evaluation of the EI is composed of :

- a control of the presence and involvement of the different students in the group's work
- a defense of the final solution in front of engineers and product managers of the partner during which the deliverables will be evaluated
- the delivery of clean, executable and commented Python code.

Course support, bibliography

+ Data Science : fondamentaux et études de cas - Machine Learning avec Python et R. E. Biernat, M. Lutz - Eyrolles

+ Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze (Cambridge University Press, 2008).
+ Massih-Reza Amini, Gaussier Eric. Recherche d'Information - applications, modèles et algorithmes. Eyrolles. Eyrolles, pp.1-233, 2013, Algorithmes, Muriel Shan Sei Fan, 978-2-212-13532-9
+ Python Data Science Handbook : https://jakevdp.github.io/PythonDataScienceHandbook/

Resources

Teaching team : Céline Hudelot and Myriam Tami and several partner employees.
Software tools :

1. Python and its data science libraries: numpy, pandas, scikit-learn, nltk, spacy,...
2. IDE left to the students' choice.
3. Git and code sharing tool (Visual Studio Code with live share option , google colab, Kaggle)
4. Group messaging : slack and Mteams

Learning outcomes covered on the course

At the end of this EI, the student will be able to:

- Apply and use a set of knowledge and information processing methods to answer and propose a solution to a real problem.

- Take a step back on an information processing problem in a real context.
- Work as a team independently and interdependently towards a common team goal.
- Defend and convince a jury of professionals.

Description of the skills acquired at the end of the course

- Applying and using a set of knowledge and information processing methods to solve a real problem is part of ***C1.1 "Studying a problem as a whole", C1.5 "Mobilizing a broad scientific and technical base as part of a transdisciplinary approach", C6.4 "Solving problems in a computational thinking approach" and C6.5 "Exploiting all types of data".***
- Taking a step back from information processing work is part of ***C4.1 "Customer Thinking", C6.6 "Understanding the Digital Economy" and C9.4 "Demonstrating rigour and critical thinking".***
- Working as a team in an autonomous and interdependent way towards a common objective for the team is part of ***C8 "Leading a project, a team".***
- Defending and convincing a jury of professionals is part of ***C7 "Knowing how to convince".***