Chapitre 4

PROJET MACHINE LEARNING

4.1 Généralité

— Les données support des projets proviennent principalement de KAGGLE
— Les projets de machine learning mènent aux thématiques suivantes :
  — classification biclasse et multiclasse
  — text mining
— On peut appliquer des méthodes supervisées et non-supervisées dans le cadre de projet de text mining
— Vous pouvez utiliser la plateforme google colab pour faire votre projet.

4.2 Premier travail

De manière synthétique vous devez :

— décrire le problème que vous devez résoudre
— définir les features et la variable cible que vous voulez prédire lors d'application de méthode supervisé
— choisire une métriques pour juger de l'efficacité de vos algorithmes dans le cadre supervisé
— dire si vous avez l'intention d'utiliser des algorithme non supervisé

4.3 Travail à rendre livrable

— Dans le cadre des ces projets vous devez rendre un notebook avec les codes qui vous ont permis de faire vos modèles de machine learning.
— Vous devez également faire une présentation power point ou bien latex beamer ayant 12 slides maximum.

4.4 Spécificitées des projets de text-mining

— Le principe du text mining est de transformer des phrases en matrices de mots, ou bien en vecteur (cas du Word embedding). Ensuite à partir de ces matrices/vecteurs on applique des modèles de machine learning.
— Les variables cibles peuvent être :

— le sujet d'un mail (topic mining)
— la note d'un produit (sentiment analysis)

Pour les groupes ayants un projet de text-mining, vous devez vous renseigner sur les Matrice termesdocuments. Vous pourrez allez sur les sites suivants :

— https://fr.wikipedia.org/wiki/Word_embedding
— https://fr.wikipedia.org/wiki/TF-IDF
— https://en.wikipedia.org/wiki/Document-term_matrix
— https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

4.5 Les projets

4.5.1 Projet 1 : Bank Marketing

— Le site kaggle correspondant à ces données est : https://www.kaggle.com/sonujha090/bank-marketing.
— Le site UCI correspondant à ces données est : https://archive.ics.uci.edu/ml/datasets/bank+marketing
— Vous pouvez aussi charger les données : https://1drv.ms/u/s!Am09h0q2OIX0ctocEPd7aZU3XIA?e=QW5HxT

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

4.5.2 Projet 2 : Bank_Loan_modelling

— Le site kaggle correspondant est :https://www.kaggle.com/itsmesunil/bank-loan-modelling
— Vous pouvez aussi charger les données : https://1drv.ms/u/s!Am09h0q2OIX0c4XP6LKqfQ3MHzU?e=O8GGlo

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns to better target marketing to increase the success ratio with a minimal budget.

The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

4.5.3 Projet 3 : Amazon Fine Food Reviews

— le site kaggle correspondant est : https://www.kaggle.com/snap/amazon-fine-food-reviews
— Vous pouvez charger les données : https://1drv.ms/u/s!Am09h0q2OIX0dHn2K8uFFzs4c5U?e=Xde9m8

This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

4.5.4 Projet 4 : Women's E-Commerce Clothing Reviews

— Le site kaggle corespondant est : https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews
— Vous pouvez charger les données : https://1drv.ms/u/s!Am09h0q2OIX0bgPKiaFTTogNLxY?e=MsUrKp

Welcome. This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions. Because this is real commercial data, it has been anonymized, and references to the company in the review text and body have been replaced with "retailer".

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables :

— Clothing ID : Integer Categorical variable that refers to the specific piece being reviewed.
— Age : Positive Integer variable of the reviewers age.
— Title : String variable for the title of the review.
— Review Text : String variable for the review body.
— Rating : Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
— Recommended IND : Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
— Positive Feedback Count : Positive Integer documenting the number of other customers who found this review positive.
— Division Name : Categorical name of the product high level division.
— Department Name : Categorical name of the product department name.
— Class Name : Categorical name of the product class name.