UMONS

Faculté des Sciences

Big Data Analytics I

Rapport de projet

Sam Boosko
Rémy Decocq
Dimitri Waelkens

Année Académique 2018-2019
Master en Sciences Informatiques
Faculté des Sciences, Université de Mons

1 Introduction

La jeu de données fourni a été construit lors de campagnes marketing menées par un organisme bancaire, sous la forme d'appels téléphoniques vers de potentiels clients. Pour chaque personne sondée, il est renseigné si oui ou non, à la suite de cet appel, elle a souscrit à un dépôt bancaire à long terme dans ladite banque. Le but de la compétition est de prédire si ce sera le cas pour de nouveau client en se basant sur des variables mesurées identiques. Le jeu d'entrainement contient 30436 observations (dont le résultat "a souscrit" est connu et est repris par la variable y), tandis ce que le jeu de test (dont le y nous est sciemment pas communiqué) est séparé en deux et contient au total $10182 \times 2 = 20364$ mesures.

Les données sont des mesures de variables qui sont de deux types : celles relatives au client lui-même (données personnelles) et celles relatives aux potentiels sondages sur le client durant les campagnes. On a donc :

Variables explicatives

`Persos :`

- *age*, (type de) *job*, statut civil *marital*, niveau du milieu d'éducation *edu*
- *default* a une défaillance de crédit, *housing* sous prêt immobilier, *loan* a un prêt personnel

`Campagnes :`

- *contact*, *month*, *day_of_week* : type de communication, mois et jour de la semaine du dernier contact
- *campaign* : nombre de contacts établis durant la campagne correspondant au jeu de données
- *pdays* : nombre de jours passés depuis le dernier contact d'une campagne précédente
- *previous* : nombre de contacts déjà établis avant cette campagne
- *poutcome* : résultat pour ce client suite à la campagne précédente

Variable expliquée

- *y* : le client a ouvert un dépôt à long terme dans l'organisme bancaire

2 Méthodologie

2.1 Analyse des variables et observations

2.2 Sélection du type de modèle

2.3 Sélection des prédicteurs pertinents

2.4 Stabilité du modèle

3 Résultats et discussion

4 Conclusion