Proposition du projet

Yassine et Mohamed

Avril 2017

1 Sujet du projet

Notre jeu de données provient du site Kaggle. Il s'agit de données sur 300 000 rendez-vous médicaux. Dans 1/3 des cas, le patient ne s'est pas présenté au rendez-vous.

Le but est de comprendre pourquoi pour ainsi prédire si un patient viendra ou pas à son rendez-vous. Lien du site : `https://www.kaggle.com/joniarroba/noshowappointments`

2 le jeu de données

Il s'agit de 300 000 rendez-vous médicaux.

Nous avons 15 caractéristiques (quantitatives et qualitatives) : l'âge, le sexe, le jour de la réservation ,le jour du rendez-vous,le jour de la semaine, le status (le patient est venu ou rendez vous ou pas), plusieurs caractéristiques sur les maladies du patient (Diabetes, Alcoolism, HiperTension, Handcap, Smokes,Tuberculosis) , est-ce que le patient bénéficie d'une aide financière, est-ce que l'on a envoyé un sms de rappel et le nombre de jours entre la prise de rendez vous et le rendez-vous lui même(un nombre négatif).

3 les méthodes à mettre en oeuvre

Après une analyse descriptive pour comprendre la structure des données et détecter les problèmes qui peuvent intervenir dans notre modélisation, notre objectif sera de mettre en oeuvre un modèle prédictif pour prédire si un patient vient au rendez-vous ($Y = 1$) ou non (Y = 0). La variable à expliquer Y est une variable binaire ou une variable quantitative avec deux modalités. Y suit une distribution de Bernoulli donc notre problème est un cas particulier de classification supervisé avec deux classes. L'une des méthodes utilisées est la régression logistique. Le problème majeur sera probablement de choisir parmi les 15 variables explicatives quelles caractéristiques nous allons utiliser dans notre modèle.