DAUPHINE
UNIVERSITÉ PARIS

Modèles linéaires et généralisés
Applications avec SAS

Projet 2018-2019

Vincent GALLMANN

Instructions

Ce projet est à réaliser à **2, 3 ou 4** personnes **maximum** avec le logiciel **SAS**.

Vous aurez le choix entre deux sujets : Une régression linéaire ou une régression logistique.

La notation prendra en compte :

- La **rédaction** d'un rapport (démarches et résultats à détailler)
- Les **résultats** (challenge sur la performance du modèle retenu)
- Le **script SAS** (suivi des recommandations faites en cours)

La date limite de rendue est le **mardi 15 janvier 2019** à 23h59.
Chaque jour de retard entraînera le retrait de 2 points.

Le tout est à envoyer à l'adresse suivante : *vincent.gallmann@dauphine.psl.eu*

Remarques

Le but ici est vraiment de vous mettre en **situation** sur un **sujet type** que vous pourrez rencontrer en **entreprise**.
Il vous est donc demandé de bien **expliciter** vos **choix**, votre **démarche**, vos **résultats**.
L'aspect **esthétique** de votre rapport + résultats sera **évalué**, notamment vos choix de représentations graphiques.

Le nombre de pages du rapport n'est pas imposé. Cependant, ce dernier ne doit ni correspondre à celui d'une nouvelle, ni à celui d'un roman.
Pas la peine de faire des pavés sur la théorie des méthodologies utilisées. Ce n'est pas ce qui est demandé.

Prédiction du prix de la drogue

Objectif

L'objectif ici sera d'expliquer le prix de différentes drogues à partir de leurs caractéristiques ainsi que de leur mode de vente.

Présentation des données

Pour réaliser cette étude, vous aurez à disposition une base *drogue.csv*

Ces données ont été relevées par des experts lors de saisies en Californie.
Chaque ligne représente une transaction.

Variable	Libellé
AN	Année de la saisie
LIEU	Endroit de la saisie : CA1 : Northeastern California, CA2 : San Francisco and the northern coastal region CA3 : Central coast CA4 : East central California CA5 : San Luis Obispo to Los Angeles CA6 : San Bernadino, Orange and Riverside counties CA7 : San Diego and Imperial counties AK : Alaska HI : Hawai OR : Oregon WA : Washington

DROGUE	Type de drogue : Coc : Cocaïne Crack : Crack Meth : Methamphetamine HerB : Heroïne Brown HerW : Heroïne White HerT : Heroïne Tar HerO : Heroïne ? MJImp : Marijuana Importée MJDom : Marijuana Domestique MJSin : Marijuana Sinsemilla Hash : Hashish
POIDS	Poids de de la drogue vendue lors de la transaction (en g)
PURETE	Pureté de la drogue (en %)
PRIX	Prix de vente (en $)

Prédiction de l'endormissement client

Objectif

L'objectif ici sera de construire un modèle prédictif de risque d'endormissement (=inactivité) des clients ayant eu une activité sur l'année 2006, voire sur l'année 2005 et qui n'ont pas acheté en 2017.

Présentation des données

Pour réaliser cette étude, vous aurez à disposition :

- Une base d'apprentissage : *achats.txt*

Ces données proviennent d'un site internet.
Vous disposerez de caractéristiques démographiques et comportementales sur 5 238 clients ayant adhérés au site Internet au cours du 2nd semestre de l'année 2004.

Les caractéristiques collectées sont les suivantes :

Variable	Libellé
IDENT	Identifiant du client
DT_ADHESION	Date d'adhésion du client
NB_CART_AAAAMM	Nb d'articles commandés sur l'année AAAA et le mois M
MONTANT_AAAA_MM	Montant dépensé sur l'année AAAA et le mois M
AGE	Age du client
SEX	Genre du client : 1 pour femme, 0 pour homme
IDF	Habite en Ile De France (0/1)

En amont vous devrez :

>> Construire la variable cible de la manière suivante :

1 si le client n'a réalisé aucun achat sur les 4 premiers mois de 2007 (mais actif en 2006)
0 sinon

>> Identifier les variables explicatives candidates (existantes ou à construire)