Exercice 1 – Scoring sous Python

PYTHON + SCIKIT-LEARN

La base « **dataset_scoring_bank_subset.xlsx** » décrit la réaction (**OBJECTIVE**), positive ou non, d'un ensemble de clients suite à une sollicitation marketing. Les variables explicatives (p01rcy...gender3) sont de natures différentes, elles peuvent décrire la récence, la fréquence et des données de type monétaire pour les comptes spécifiques, ou encore représenter des variables de recensement ou démographiques. Pour nous simplifier la tâche, on considère qu'une normalisation des données n'est pas nécessaire dans cet exercice.

La colonne **ExStatus** est particulière, elle permet de scinder les données en échantillons d'apprentissage (TRAIN) et de test (TEST).

Le but de l'étude est de cibler les individus qui réagissent positivement à la sollicitation commerciale (OBJECTIVE = POSITIVE).

Nos tutoriels de référence sont :

a. Manipulation de données avec Pandas : http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#594407616861142837 [**TUTO 1**]

b. Machine Learning avec scikit-learn : http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#1766550852310168124 [**TUTO 2**]

Importation et préparation des données

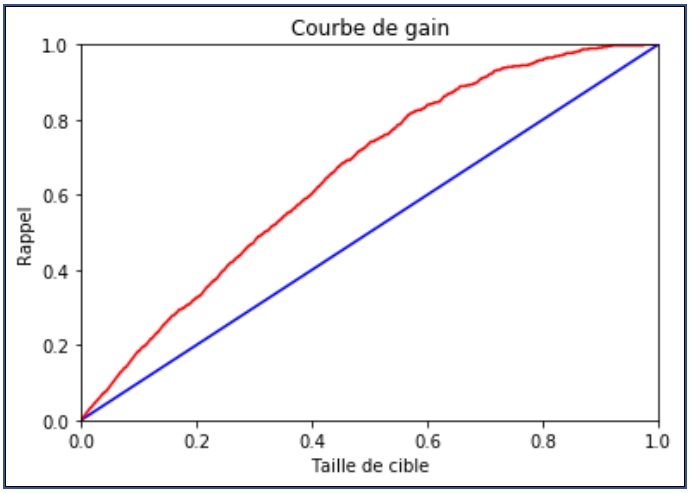
1. Importez les données de « **dataset_scoring_bank_subset.xlsx** » dans un DataFrame.
2. Affichez les premières lignes (head). Combien a-t-il d'observations et de variables ? (shape) (2158, 17).
3. Affichez la liste des noms de variables (columns.values).
4. Scindez les données en échantillons d'apprentissage et de test **en vous appuyant sur la colonne ExStatus** [**TUTO 1, page 9 ;** voir aussi https://youtu.be/gweedJd3cvw?t=2567]. Attention, la colonne **ExStatus** ne sera plus utilisée par la suite, il faut l'évacuer. Quelles sont les dimensions des DataFrame obtenus ? [train : (1158, 16) ; test : (1000, 16)].
5. Calculez les proportions d'observations positives et négatives dans les deux sous-échantillons (value_counts, voir http://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.value_counts.html) (≈ 0.51 et 0.49).
6. Pour les deux échantillons, préparez les structures matricielles en mettant les X (variables explicatives) d'un côté, la colonne y (variable cible, objective) de l'autre [**TUTO 2, page 8**].

Modélisation avec l'analyse discriminante

7. Importez la méthode **`LinearDiscriminantAnalysis`** de scikit-learn. Cf. https://www.youtube.com/watch?v=yoXZztiqULM pour l'importation et l'instanciation d'une classe de calcul. Nous conservons les paramètres par défaut.
8. Construisez le modèle sur les données d'apprentissage (fit).
9. Affichez les coefficients (.coef_) et la constante (.intercept_) de la fonction de classement.

Construction de la courbe de gain

10. Calculez et affichez les probabilités d'affectation sur l'échantillon test (predict_proba).
11. Pour identifier dans quelle colonne est située la modalité positive (objective = positive), affichez la liste des classes du modèle (classes_). Récupérez alors le score c.-à-d. la probabilité d'être positif dans un vecteur spécifique. Calculez la moyenne des scores à titre de vérification (0.524227...).
12. Construisez et dessinez la courbe de gain en vous inspirant du **TUTO 2, pages 17 à 19**.



Interprétation de la courbe

13. Si l'on applique la fonction score sur une base marketing comportant 50.000 observations. Combien d'observations positives obtiendrait-on si nous sélectionnons les 15.000 individus qui présentent les scores les plus élevés ? (11600).