Data Mining
Projet

Consignes

- **Le travail est à réaliser en binôme.**
- **Les noms des étudiants binômés sont à envoyer à julien.blanchard@univ-nantes.fr avant le 28 avril 2018.**
- **Toute oeuvre de plagiat entre binômes sera sanctionnée par une note nulle (plagieurs comme plagiés).**
- **Toute oeuvre de plagiat depuis internet sera sanctionnée par une note nulle.**
- **Date limite de remise des travaux (sources+rapports) sur Madoc dans la section "Data Mining (X2II020)" : 18 mai 2018**

1 Données

Les données proviennent d'*Alphaprise*, une entreprise B2B qui vend des produits et services. Elles décrivent les clients de l'entreprise et en particulier les chiffres d'affaires réalisés dans différentes familles de produits.

Les 11566 clients sont des entreprises de différents secteurs d'activité et de différentes tailles (TPE, PME, ETI). Elles sont décrites par les variables suivantes :

- **Client** : l'identifiant de l'entreprise cliente
- **NbSalaries** : nombre de salariés de l'entreprise cliente
- **CapaciteEmprunt** : montant annuel maximum autorisé pour un prêt d'*Alphaprise* envers son client
- **PrévisionnelAnnuel** : chiffre d'affaires annuel prévisionnel engendré par ce client pour *Alphaprise*
- **P1** à **P30** : les chiffres d'affaires engendrés par ce client sur 30 familles de produits sur l'année écoulée (des redondances peuvent exister entre les variables P1 à P30)

ainsi que d'autres variables issues de la direction des ventes, ajoutées pour servir de cibles aux modèles d'apprentissage supervisé de la partie 2 :

- **Secteur1** : indique si le client relève du secteur d'activité codé 1
- **Secteur2** : indique si le client relève du secteur d'activité codé 2
- **SecteurParticulier** : indique si le client est en fait un particulier.

Le nombre de salariés donne une bonne idée de la taille de l'entreprise. Plus l'entreprise a de salariés, plus elle devrait engendrer un chiffres d'affaires important pour Alphaprise.

Les variables PrévisionnelAnnuel et CapaciteEmprunt sont des variables externes achetées auprès d'un fournisseur de données.

L'entreprise vise à terme deux objectifs :

— s'affranchir du fournisseur de données en estimant elle-même les variables PrévisionnelAnnuel et CapaciteEmprunt ;
— déduire des données la notion de secteur d'activité, pour ainsi prendre en compte les changements de secteurs de ses clients.

2 Travail demandé

— Construire un modèle pour estimer la variable CapacitéEmprunt à partir des données, puis un modèle pour la variable PrévisionnelAnnuel.
Variables prédictives : P1 à P30 et le nombre de salariés.

— Construire un modèle de scoring pour estimer Secteur1, puis Secteur2, puis SecteurParticulier.
Variables prédictives : P1 à P30 et le nombre de salariés.
La direction des ventes fournit l'indication suivante : si deux clients ont le même secteur d'activité, les proportions de chiffre d'affaires sur chaque famille Pi sont proches chez les deux clients (proportion par rapport au chiffre d'affaires total que chaque client réalise avec *Alphaprise*).

Pour chaque tâche, vous devez décrire les pré-traitements réalisés pour préparer les données, et évaluer/valider vos résultats.

Un ensemble de test vous est également fourni. Il contient 1000 clients. A l'aide des modèles que vous aurez construits, il faudra renseigner vos prédictions dans les dernières colonnes du tableau (en jaune).

Conseil : documentez bien tous les pré-traitements réalisés sur les données d'apprentissage, car vous aurez à les refaire à l'identique sur les données de test.

3 A remettre sur Madoc

— Vos sources : programmes et scripts, fichiers de projets si vous avez utilisé un logiciel dédié.
— Un rapport PDF (10 à 15 pages) qui présente le travail réalisé, avec en particulier :
  — les divers pré-traitements réalisés pour préparer les données,
  — les phases d'apprentissage (choix de l'algorithme, réglage des hyperparamètres),
  — l'évaluation des modèles,
— Les prédictions sur le jeu de test.