Analyse et Fouille de Données
Mini-Projet ID5

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
- **Activité** : indique si le client est actif ou inactif depuis un an
- **CodeGroupe** et **Programme** : informations commerciales
- **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*
- **NbSalaries** : nombre de salariés de l'entreprise cliente
- **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
- **SecteurDivers** : indique que le client relève du secteur d'activité "divers".

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 deux objectifs :

- améliorer la qualité de la variable "secteur d'activité" en la déduisant directement des données, pour ainsi prendre en compte les changements de secteurs de ses clients ;
- s'affranchir du fournisseur de données en estimant elle-même les variables PrévisionnelAnnuel et CapaciteEmprunt.

2 Travail demandé

- Réaliser une analyse statistique de base.
- Faire émerger des secteurs d'activité à partir des données (le nombre de secteurs d'activité identifiés par la direction des ventes est de 8 secteurs, en comptant les Particuliers et les Divers, mais ce chiffre n'est en aucun cas une valeur parfaite).
  Variables à exclure : au moins Secteur1, Secteur2, SecteurParticulier, SecteurDivers.
- Construire une représentation graphique des secteurs d'activités (type cartographie).
  Variables illustratives : au moins Secteur1, Secteur2, SecteurParticulier, SecteurDivers.
- Construire un modèle pour estimer la variable CapacitéEmprunt à partir des données, puis un modèle pour la variable PrévisionnelAnnuel.
  Variables à exclure : au moins Secteur1, Secteur2, SecteurParticulier, SecteurDivers, CapacitéEmprunt, PrévisionnelAnnuel.
- Construire un modèle de scoring pour estimer Secteur1, puis Secteur2, puis SecteurParticulier.
  Variables à exclure : toutes sauf P1 à P30 et à la rigueur le nombre de salariés.

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

- Vos sources : programmes et scripts, fichiers de projets si vous avez utilisé un logiciel dédié.
- Un rapport PDF (15 à 20 pages) qui présente le travail réalisé, avec en particulier :
  - les divers pré-traitements réalisés pour préparer les données,
  - les phases d'analyse (sélection des algorithmes, réglage des hyperparamètres),
  - l'évaluation des résultats,
  - et pour les modèles prédictifs, les prédictions sur le jeu de test.

4 Evaluation

- Le travail est à réaliser en binôme.
- Chaque binôme dépose ses sources et son rapport sur Madoc avant le **11 février**.
- Chaque binôme présente son travail sur machine en 10 à 15 minutes le 12 février en D009.