Projet : Prédiction de l'Appétence de Clients

La définition d'offres promotionnelles commerciales efficaces nécessite la compréhension des différents groupes, ou segments, de clients, afin de mieux cibler ces derniers et ainsi à la fois maximiser le taux de réponses positives et limiter les coûts de diffusion. L'objectif de cette étude est l'analyse d'un ensemble de données contenant les réponses à un offre promotionnelle commerciale afin de comprendre quels sont les critères associés à une réponse positive ou négative par les clients et de construire un modèle de prédiction de l'appétence de clients, c-à-d leur propension à accepter ou refuser l'offre promotionnelle.

Ce projet peut être réalisé en monôme ou en binôme. Si pour une raison justifiée (e.g., nombre impair d'étudiants) vous souhaitez le réaliser en trinôme, contactez-moi par e-mail afin que je puisse valider votre demande.

1. Ensembles de Données

L'ensemble de données utilisé pour cette étude est issu d'une base de données de clientèle acquise en vue de la diffusion d'offres promotionnelles.

Chacune des 6400 instances de l'ensemble de données représente un client pour lequel on dispose :

- de diverses informations socio-démographiques (âge, etc.) décrites par 28 variables hétérogènes,
- de l'indication si le client a répondu positivement ou non à l'offre par la variable RESPONSE.

Les caractéristiques des variables sont décrites dans le dictionnaire des données ci-dessous.

Dictionnaire des Données

VARIABLE	TYPE	DESCRIPTION	VALEURS
AGE	Numérique	Age en années	[18, 77]
MARITAL	Booléen	Statut marital	0 = non marié, 1 = marié
ADDRESS	Numérique	Nombre d'années à l'adresse actuelle	[0, 56]
INCOME	Numérique	Revenu du foyer en milliers ($)	[9, 1116]
INCCAT	Ordinal	Catégories de revenu en milliers ($)	1 = [0, 25[, 2 = [25, 50[, 3 = [50, 74[, 4 = [75, +?[
CAR	Numérique	Prix du premier véhicule en milliers ($)	
CARCAT	Ordinal	Catégorie de prix du véhicule premier	1 = économique, 2 = standard, 3 = luxueux
ED	Ordinal	Nombre d'années d'éducation	1 = inf. au bac, 2 = niveau bac, 3 = bac + 2, 4 = bac + 3/4, 5 = bac +5 et +
EMPLOY	Numérique	Nombre d'années chez l'employeur actuel	[0, 57]
RETIRE	Booléen	Retraité(e)	0 = non, 1 = oui
EMPCAT	Ordinal	Nombre d'années chez l'employeur actuel (catégorie)	0 = [0, 5[, 1 = [5, 15[, 2 = [15, +?[
JOBSAT	Ordinal	Satisfaction Emploi	1 = très insatisfait, 2 = insatisfait, 3 = neutre, 4 = satisfait, 5 = très satisfait
GENDER	Nominal	Sexe	F = femelle, M = mâle
RESIDE	Ordinal	Nombre de personnes dans le foyer	[1, 9]
WIRELESS	Booléen	Service sans fil	0 = non, 1 = oui
MULTLINE	Booléen	Lignes multiples	0 = non, 1 = oui
VOICE	Booléen	Boîte vocale	0 = non, 1 = oui
PAGER	Booléen	Service Paging	0 = non, 1 = oui
INTERNET	Catégoriel	Internet	0 = non, 1 = oui, 8 = ne sais pas, 9 = pas de réponse
CALLID	Booléen	Affichage numéro d'appel	0 = non, 1 = oui
CALLWAIT	Booléen	Double appel	0 = non, 1 = oui

VARIABLE	TYPE	DESCRIPTION	VALEURS
OWNTV	Booléen	Possesseur d'un téléviseur	0 = non, 1 = oui
OWNVCR	Booléen	Possesseur d'un magnétoscope	0 = non, 1 = oui
OWNCD	Booléen	Possesseur de matériel HiFi	0 = non, 1 = oui
OWNPDA	Booléen	Possesseur d'un agenda électronique	0 = non, 1 = oui
OWNPC	Booléen	Possesseur d'un ordinateur	0 = non, 1 = oui
OWNFAX	Booléen	Possesseur d'un fax	0 = non, 1 = oui
NEWS	Booléen	Abonnement à un magasine	0 = non, 1 = oui
RESPONSE	Booléen	Réponse positive à l'offre	0 = non, 1 = oui

Les caractéristiques des deux fichiers de données fournis pour réaliser cette étude sont décrites dans la table ci-dessous.

Fichiers de données

Fichier	Nbr instances	Classe?	Remarques
Data Projet.csv	6400	Oui	Instances dont la classe réelle est connue
Data Projet New.csv	300	Non	Instances à prédire

2. Objectifs du projet

L'objectif est d'identifier les critères (caractéristiques des clients) principaux associés à une réponse positive ou négative des clients à l'offre et de créer un modèle de prédiction de la réponse des clients à l'offre pour l'appliquer ce modèle à un ensemble de nouveaux clients (instances à prédire). On souhaite donc utiliser les techniques de clustering afin de générer des clusters distinguant les deux classes de clients (réponse positive et réponse négative) et les techniques de classification supervisée afin de générer un modèle de prédiction de la classe des clients :

- RESPONSE = 1 (réponse positive)
- RESPONSE = 0 (réponse négative)

Clustering

L'utilisation des techniques de clustering a pour objectif d'identifier des groupes de clients qui appartiennent à la même classe (réponse positive ou réponse négative) et qui ont des caractéristiques (valeurs des autres variables) semblables ou proches. Le but est de générer des clusters qui correspondent chacun autant que possible à une classe unique, plusieurs clusters pouvant correspondre à la même classe (si plusieurs groupes de clients correspondent à la même classe), et d'identifier les caractéristiques communes (moyenne pour les variables numériques et valeur la plus fréquente pour les variables discrètes) aux instances de chaque cluster. Parmi ces caractéristiques communes, celles permettant de distinguer les clients appartenant à différents clusters seront représentatives des caractéristiques distinctives des clients répondant positivement et négativement à l'offre.

Classification Supervisée

Plusieurs classifieurs seront générés et testés en appliquant les différentes méthodes de classification et en ajustant les paramètres afin d'optimiser les résultats.
Seul le classifieur le plus performant sera conservé sachant que l'on souhaite avant tout minimiser le risque de ne pas prévoir l'appétence d'un client, c'est-à-dire que l'on souhaite avant tout éviter de prédire à tort un client comme non susceptible de répondre positivement à l'offre.
Le classifieur sélectionné sera ensuite appliqué à l'ensemble de données à prédire afin de prédire pour chaque client s'il est susceptible de répondre positivement (classe RESPONSE = oui) ou non (classe RESPONSE = non) son contrat.
Afin d'évaluer les classifieurs générés, vous définirez un ou des critère(s) (basés sur les taux de succès/échecs, la matrice de confusion ou les mesures d'évaluation par exemple) en fonction des critères de sélection du classifieur décrits précédemment. Vous comparerez les résultats des classifieurs générés selon ce(s) critère(s) afin d'identifier le plus pertinent.

3. Processus d'analyse

Le processus général pour cette analyse suivra les étapes suivantes :

- Exploration et visualisation des données.
- Pré-traitements éventuels des données.
- Clustering des données.
- Classification supervisée :
    - Définition de la méthode d'évaluation des classifieurs.
    - Définition des données d'apprentissage et de test.
    - Construction et évaluation des classifieurs.
    - Choix du classifieur le plus performant.
    - Application du classifieur aux données à prédire.

Référez-vous aux méthodes appliquées durant les séances de Travaux Dirigés pour chacune de ces étapes.

4. Rapport de projet

Vous devez déposer votre rapport de projet sur le LMS UCA sous forme d'un unique fichier archive (format zip, 7z, rar, tgz, etc.) à la **date indiquée sur l'onglet du projet**.
Ce fichier archive doit comporter uniquement les trois fichiers suivants :

- Un rapport au **format .pdf** décrivant tous les traitements que vous avez effectué et les résultats obtenus :
    - Indiquez votre(vos) **nom(s) et prénom(s) sur la première page** du rapport.
    - Exploration des données et interprétation des résultats (relations notables, problèmes, variables ou valeurs les plus utiles pour la prédiction de la classe, associations, etc.).
    - Pré-traitements appliqués aux données si besoin (sélection des variables, transformation des valeurs, etc.).
    - Méthode de clustering utilisée et résultats obtenus correspondant aux différents clusters éventuels pour chacune des classes, et caractéristiques spécifiques des instances de chacun de ces clusters, c-à-d qui les distinguent de celles des autres clusters.
    - Définition de la méthode d'évaluation des classifieurs (ex : taux de succès/échecs, matrices de confusion, mesures d'évaluation, courbe ROC, etc.) pour la sélection du classifieur le plus pertinent en fonction des objectifs de l'application décrits précédemment.
    - Description de la méthode de création des données d'apprentissage et de test : techniques utilisées (partitionnement, échantillonnage, etc.) et leur paramétrage(s), etc.
    - Description des configurations des classifieurs générés (algorithmes et paramétrages) et évaluation de leur performances selon la méthode d'évaluation définie précédemment. Vous indiquerez quel(s) est(sont) le(s) classifieur(s) donnant les meilleurs résultats selon cette méthode d'évaluation.
    - Description du classifieur sélectionné (type de modèle généré, algorithme et paramétrage utilisé, etc.) et de sa structure en fonction du type de classifieur et des options utilisées (ex : dimensions de l'arbre de décision, nombre de règles de classification, nombre et tailles des couches du réseau de neurones, etc.). C'est à dire tous les éléments qui vous paraissent utiles pour décrire sa structure, sa complexité et sa pertinence (autres mesures, courbe ROC, etc.).
    - Résumé des résultats de l'application du classifieur sélectionné à l'ensemble de données à prédire (distribution des classes prédites, probabilités minimales, maximales et moyennes associées à chacune des classes, etc.).
    - Conclusion résumant vos autres observations sur cette application et les résultats, les difficultés rencontrées, etc.
- Un fichier au **format .csv** contenant les résultats de l'application du classifieur sélectionné à l'ensemble à prédire afin de fournir une prédiction de la classe pour chacune des instances à prédire. Le résultat doit être représenté sous forme d'un tableau avec sur chaque ligne :
    - Le numéro d'identification de l'instance
    - La classe prédite pour cette instance.
    - La probabilité associée à la prédiction de cette classe.
- Un fichier au **format .R** contenant le script R commenté (indiquez pour chaque groupe de lignes l'opération réalisée) des commandes R utilisées pour réaliser le projet.

Consigne : le fichier archive déposé doit porter les **noms et prénoms des auteurs**, par exemple *PASQUIER_Nicolas_DUPOND_Jean*.zip.