Prédiction de Défauts de Paiement

Ce projet peut être réalisé en monôme ou en binôme. Vous utiliserez pour le réaliser les ensembles de données `Data Projet.csv` et `Data Projet New.csv` disponibles sur la page du cours.

Ensembles de données

Ces deux ensembles de données concernent les mesures qu'entreprend une banque pour réduire le taux de défauts de paiement des remboursements d'emprunts.

Le fichier `Data Projet.csv` contient des informations financières et démographiques concernant 1200 clients ayant déjà effectué un emprunt, avec pour chacun l'information sur un défaut de paiement survenu ou non (variable `default`).

Le fichier `Data Projet New.csv` contient les informations sur 300 clients pour lesquels la banque souhaite prédire s'il y a un risque de défaut de paiement pour l'octroi d'un emprunt.

Caractéristiques des données :

- Instances : chaque instance correspond à un client identifié par son numéro
- Nombre de variables : 12
- Séparateur de colonnes : virgule
- Séparateur de décimales : point
- Variable de classe : `default`
- Valeurs manquantes : aucune

Le dictionnaire des données ci-dessous décrit pour chacune des 12 variables son nom, son type (entier, réel, booléen, catégoriel ou ordinal), sa description et son domaine de valeurs (liste de valeurs ou nombres minimal et maximal).

Dictionnaire des données

Variable	Type	Description	Domaine de valeurs
branch	Catégoriel	Code de la branche d'activité du client.	{3, 13, 15, 20, 25, 49, 60, 64, 68, 73, 74, 75, 76, 77, 91}
ncust	Entier	Nombre de clients dans la branche d'activité.	[1919, 4809]
customer	Entier	Numéro unique d'identification du client	[10010, 453800]
age	Entier	Age en nombre d'années	[18, 79]
ed	Catégoriel	Niveau d'éducation relativement au baccalauréat	{Niveau bac, Bac+2, Bac+3, Bac+4, Bac+5 et plus}
employ	Entier	Nombre d'années avec l'employeur actuel	[0, 63]
address	Entier	Nombre d'années à l'adresse actuelle	[0, 34]
income	Réel	Revenus du foyer en milliers de $	[12.0, 1079.0]
debtinc	Réel	Débit carte de crédit en milliers de $	[0.0, 40.7]
creddebt	Réel	Ratio Débit/Crédit (x100)	[0.00, 35.97]
othdebt	Réel	Autres dettes en milliers de $	[0.00, 63.47]
default	Booléen	Un défaut de paiement a-t-il eu lieu ? Variable de classe.	{Oui, Non}

Fichiers de données

Fichier	Nbr instances	Classe?	Remarques
`Data Projet.csv`	1200	Oui	Instances dont la classe réelle est connue
`Data Projet New.csv`	300	Non	Instances à prédire

Objectifs du projet

L'objectif est la création d'un modèle de prédiction du risque de défaut de paiement pour les clients et son application aux instances à prédire. On souhaite donc utiliser les techniques de classification afin de générer un modèle de prédiction de la classe des clients :

- `default = Oui` (positif)
- `default = Non` (négatif)

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 les risques financiers en évitant d'accorder un emprunt à tort, c-à-d d'accorder un emprunt à un client pour lequel un défaut de paiement est prévisible.

Le classifieur sélectionné sera ensuite appliqué à l'ensemble de données à prédire afin de prédire pour chaque client s'il est susceptible d'avoir un défaut de paiement (classe `default = Oui`) ou non (classe `default = Non`).

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 objectifs de l'application décrits ci-dessus. Vous comparerez les résultats des classifieurs générés selon ces critères afin d'identifier le plus pertinent.

Processus d'analyse

Le processus général pour cette analyse suivra les étapes suivantes :

- Exploration et visualisation des données.
- Pré-traitement des données.
- 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.

Rapport de projet

Vous devez déposer votre rapport de projet sur la plate-forme LMS UCA **au plus tard le dimanche 25 avril 2021**. Si vous travaillez en binôme, ne faite qu'un seul dépôt par le compte de un de vous deux et indiquez vos deux noms dans le nom des fichiers (ex : *PASQUIER_Nicolas_DUPOND_Jean*.pdf).

Les instructions concernant le dépôt du projet vous sont rappelées sur l'onglet dédié du LMS UCA.

Les **trois fichiers** constituant votre rapport de projet à déposer sur le LMS UCA sont :

- 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.).
  - Définition de la méthode d'évaluation des classifieurs (taux de succès/échecs, matrices de confusion, mesures d'évaluation, etc.) pour la sélection du classifieur le plus pertinent en fonction des objectifs.
  - 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, algorithme, paramétrage, etc.) et de sa structure en fonction du type de classifieur et des options utilisées (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.
  - Résumé des résultats de l'application du classifieur sélectionné à l'ensemble de données à prédire (répartition des classes, 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 chacun des nouveaux clients.

  Le résultat doit être représenté sous forme d'un tableau avec sur chaque ligne uniquement :

  - Le numéro d'identification du client.
  - La classe prédite pour ce client.
  - La probabilité associée à la prédiction de cette classe.

- Un fichier au **format .R** contenant le script R des commandes R utilisées pour réaliser le projet. Commentez les parties les plus importantes (blocs de ligne réalisant une opération ou commande complexe par exemple) de votre fichier de code afin d'en faciliter la lecture et réutilisation.

Le non-respect des consignes entrainera une diminution de la note.