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 des 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 des 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
- Noms des variables sur la première ligne : oui
- Séparateur de colonnes : virgule
- Séparateur de décimales : point
- Variable de classe : `default`
- Valeurs manquantes : `age`, `adresse`

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

Dictionnaire des données

Variable	Type	Description	Domaine de valeurs	Valeurs manquantes
`client`	Entier	Numéro d'identification du client	`[1201, 8500]`	
`age`	Entier	Age en nombre d'années	`[18, 999]`	`999`

`education`	Ordinal	Niveau d'éducation relativement au baccalauréat	`Niveau bac`
			`Bac+2`
			`Bac+3`
			`Bac+4`
			`Bac+5 et plus`	

`emploi`	Entier	Nombre d'années avec l'employeur actuel	`[0, 63]`	
`categorie`	Entier	Catégorie bancaire	`[12, 12]`	
`adresse`	Entier	Nombre d'années à l'adresse actuelle	`[0, 999]`	`999`
`revenus`	Réel	Revenus du foyer en milliers de $	`[12.3, 2461.7]`	
`debcred`	Réel	Ratio Débit/Crédit (x100)	`[0.08, 44.62]`	
`debcarte`	Réel	Débit carte de crédit en milliers de $	`[0.005, 139.580]`	
`autres`	Réel	Autres dettes en milliers de $	`[0.009, 416.517]`	
`defaut`	Booléen	Un défaut de payement a-t-il eu lieu ?	`Oui` `Non`	

Les deux fichiers à utiliser sont décrits dans le tableau ci-dessous.

Fichiers de données

Fichier	Nbr instances	Classe?	Remarques
`Data Projet.csv`	6000	Oui	Instances dont la classe réelle est connue
`Data Projet New.csv`	500	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 :

- Pré-traitement des données.
- Analyse exploratoire 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 tutoriels pour chacune de ces étapes.

Rapport de projet

Si vous travaillez en binôme, ne transmettez qu'un seul rapport pour vous deux et nommez vos fichiers avec vos deux noms (ex : *PASQUIER_Nicolas_DUPOND_Jean*.pdf).

Les **trois fichiers** constituant votre rapport de projet 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.
  - Pré-traitements appliqués aux données si besoin (sélection des variables, typage des variables, transformation des valeurs, etc.).
  - Analyse exploratoire des données et interprétation des résultats (relations notables, problèmes, variables ou valeurs les plus utiles pour la prédiction de la classe, etc.).
  - Définition de la(des) méthode(s) 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(les) méthode(s) d'évaluation définie(s) 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.), de sa structure en fonction du type de classifieur et des options utilisées (dimensions de l'arbre de décision, nombre de règles de classification, etc.) et de ses performances détaillées (taux de succès, mesures, 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.

Consignes

- Mentionnez dans votre rapport tout ce que vous avez testé (méthodes, algorithmes, paramétrages, visualisations, comparaisons, etc.), même si cela n'a pas donné de résultat probant ou utile.

  Le fait qu'un test réalisé ne donne pas de résultat pertinent est une connaissance utile pour la compréhension de l'application, des données, etc.

  Inutile d'entrer dans le détail sur ces points toutefois, soyez concis et bref sur ceux-ci s'il y en a.

- Indiquez votre(vos) nom(s) et prénom(s) sur la première page du rapport.

- Nommez vos fichier avec les NOMS et Prénoms des membres de votre groupe.

  Par exemple :

  - PASQUIER_Nicolas_DUPOND_Jean.pdf
  - PASQUIER_Nicolas_DUPOND_Jean.csv
  - PASQUIER_Nicolas_DUPOND_Jean.R