Projet Machine Learning : Prédiction de Défauts de Paiement (KARKOURI Zakaria)

Le projet vise à développer un modèle de prédiction du risque de défaut de paiement pour les clients d'une banque. Les données utilisées comprennent des informations financières et démographiques sur 1200 clients ayant déjà effectué un emprunt, ainsi que 300 nouveaux clients pour lesquels la banque souhaite prédire le risque de défaut de paiement.

Exploration des Données

L'étape d'exploration des données a été cruciale pour assurer la qualité des informations avant l'application des techniques de classification. Voici quelques détails sur cette phase :

Vérification des Valeurs Incorrectes

Nous avons commencé par inspecter chaque variable afin d'identifier d'éventuelles valeurs aberrantes ou incorrectes. Cela implique la vérification des plages de valeurs attendues, des types de données, et la détection de toute anomalie. Aucune valeur incorrecte n'a été identifiée dans les données, ce qui indique une qualité globale des enregistrements.

Gestion des Valeurs Manquantes

Une inspection attentive a été réalisée pour détecter toute valeur manquante dans les variables. Heureusement, aucun cas de valeurs manquantes n'a été trouvé, ce qui simplifie le processus d'analyse et garantit la complétude des données.

Suppression des Variables Inutiles

Certaines variables, telles que 'customer' et 'ncust', ont été identifiées comme inutiles pour la tâche de prédiction du défaut de paiement. Ces variables n'apportent pas d'informations significatives pour la classification des clients. Par conséquent, elles ont été supprimées pour simplifier le modèle et accélérer les calculs.

Pré-traitement des Données

Suite à l'exploration des données, nous avons préparé les données en les divisant en ensembles d'apprentissage (EA) et de test (ET). Cette division est essentielle pour évaluer les performances des classifieurs sur un ensemble de données independent.

Méthode d'Évaluation des Classifieurs

Pour évaluer les classifieurs, nous avons utilisé le taux de réussite en comparant les prédictions avec les valeurs réelles. Les classifieurs ont été générés en utilisant différentes méthodes, notamment rpart, C5.0 et Tree. Les paramètres ont été ajustés pour optimiser les performances.

rpart

- **split = "gini" et minbucket = 10**
    - Paramètre de division basé sur l'indice de Gini, avec un seuil minimum de 10 observations dans une feuille.
- **split = "gini" et minbucket = 5**
    - Paramètre de division basé sur l'indice de Gini, avec un seuil minimum de 5 observations dans une feuille.
- **split = "information" et minbucket = 10**
    - Paramètre de division basé sur l'information, avec un seuil minimum de 10 observations dans une feuille.
- **split = "information" et minbucket = 5**
    - Paramètre de division basé sur l'information, avec un seuil minimum de 5 observations dans une feuille.

C5.0

- **minCases = 10 et noGlobalPruning = FALSE**
    - Minimum de 10 cas par nœud, avec l'autorisation de la taille globale du modèle.
- **minCases = 10 et noGlobalPruning = TRUE**
    - Minimum de 10 cas par nœud, sans autorisation de la taille globale du modèle.
- **minCases = 5 et noGlobalPruning = FALSE**
    - Minimum de 5 cas par nœud, avec l'autorisation de la taille globale du modèle.
- **minCases = 5 et noGlobalPruning = TRUE**
    - Minimum de 5 cas par nœud, sans autorisation de la taille globale du modèle.

tree

- **split = "deviance" et mincut = 5**

    - Paramètre de division basé sur le critère de deviance, avec un seuil minimum de 5 observations dans une feuille.
- **split = "deviance" et mincut = 10**
    - Paramètre de division basé sur le critère de deviance, avec un seuil minimum de 10 observations dans une feuille.
- **split = "gini" et mincut = 5**
    - Paramètre de division basé sur l'indice de Gini, avec un seuil minimum de 5 observations dans une feuille.
- **split = "gini" et mincut = 10**
    - Paramètre de division basé sur l'indice de Gini, avec un seuil minimum de 10 observations dans une feuille.
    -

Construction et Évaluation des Classifieurs

Plusieurs classifieurs ont été générés en utilisant différents ensembles de paramètres pour les algorithmes rpart, C5.0, et tree. Chaque configuration a été évaluée sur l'ensemble de test pour déterminer sa performance dans la prédiction du risque de défaut de paiement. Les résultats détaillés de chaque classifieur ont été enregistrés pour une analyse comparative.

Performances des Classifieurs

Les résultats d'évaluation des classifieurs sur l'ensemble de test sont résumés dans le tableau ci-dessous :

Algorithme	Paramètres	Taux de Réussite (%)
rpart	split = "gini" et minbucket = 10	76.25
rpart	split = "gini" et minbucket = 5	68.25
rpart	split = "information" et minbucket = 10	69.5
rpart	split = "information" et minbucket = 5	68.25
C5.0	minCases = 10 et noGlobalPruning = FALSE	74.5
C5.0	minCases = 10 et noGlobalPruning = TRUE	74.75
C5.0	minCases = 5 et noGlobalPruning = FALSE	75.25
C5.0	minCases = 5 et noGlobalPruning = TRUE	74.25

tree	split = "deviance" et mincut = 5	**76.5 (Optimal)**
tree	split = "deviance" et mincut = 10	76.25
tree	split = "gini" et mincut = 5	68.5
tree	split = "gini" et mincut = 10	71.25

Choix du Classifieur Optimal

Le classifieur basé sur l'algorithme Tree avec un critère de découpe 'deviance' et un seuil minimal de 5 a démontré la meilleure performance, avec un taux de réussite de 76.5%. Par conséquent, ce classifieur a été choisi comme le plus optimal pour la prédiction du risque de défaut de paiement.

Application du Classifieur aux Données à Prédire

Le classifieur optimal a été appliqué à l'ensemble de données des nouveaux clients (Data Projet New.csv) afin de prédire le risque de défaut de paiement pour chaque client. Les résultats ont été sauvegardés dans un fichier CSV contenant la classe prédite pour chaque client, ainsi que la probabilité associée à cette prédiction. Ces informations seront essentielles pour prendre des décisions éclairées sur l'octroi d'emprunts aux nouveaux clients.

Résumé des résultats

Ces résultats sont essentiels pour la banque. Il est observé que sur les 300 nouveaux clients, le modèle prédit que 224 n'auront pas de défaut de paiement, tandis que 76 présentent un risque de défaut selon le modèle.

Cette information peut être utilisée par la banque pour prendre des décisions éclairées lors de l'octroi d'emprunts. Les clients identifiés comme présentant un risque de défaut pourraient être soumis à des évaluations plus approfondies avant d'accorder un prêt, permettant ainsi à la banque de gérer plus efficacement son risque financier.

En conclusion, ce projet de prédiction de défaut de paiement a été mené avec succès en suivant une méthodologie rigoureuse. L'étape d'exploration des données a permis de garantir la qualité des informations en identifiant et traitant toute anomalie potentielle. Le pré-traitement des données, notamment la division en ensembles d'apprentissage et de test, a créé une base solide pour l'évaluation des classifieurs.

Les classifieurs générés à l'aide des algorithmes rpart, C5.0 et tree ont été minutieusement évalués, avec des ajustements de paramètres pour optimiser leurs performances. Après une analyse comparative, le classifieur basé sur l'algorithme Tree, utilisant le critère de découpe 'deviance' avec un seuil minimal de 5, a émergé comme le plus performant, atteignant un taux de réussite de 76.5% sur l'ensemble de test.

L'application de ce classifieur optimal aux nouvelles données clients a permis de prédire le risque de défaut de paiement, fournissant des informations cruciales pour la prise de décisions éclairées quant à l'octroi d'emprunts.