

Income Prediction for targeted marketing campaigns

Vous travaillez pour une entreprise de marketing qui a accès à des données démographiques. Vous avez pour mandat de construire un modèle qui permettra de prédire avec précision si une personne aura un revenu qui sera supérieur ou inférieur à 50K\$. Vous avez accès à un jeu de données qui décrit différents individus avec les attributs suivants :

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

Le *education-number* est une représentation numérique de l'attribut éducation. Les attributs *capital-gain* et *capital-loss* donnent respectivement le gain en capital et la perte de l'individu. L'attribut *fnlwgt* est un score démographique attribué à un individu sur la base d'informations telles que la zone de résidence et le type d'emploi.

Votre mission: l'agence vous a mandaté pour analyser les données afin de construire un modèle qui permettra de prédire qui sont les individus qui gagnent plus de 50K \$. Votre analyse inclura un certain nombre de différentes étapes

- 1) Exploration simple des données, calcul des statistiques de base, visualisation et inspection des données. Êtes-vous capable de trouver des attributs importants déjà à cette étape.
- 2) Explorer et évaluer différentes algorithmes d'apprentissage pour construire vos modèles prédictives.
- 3) Sur la base de l'analyse des modèles, déterminer quels sont les attributs les plus importants.
- 4) Presentez vos résultats.