

SCORING

Loic Hannequin M2 DS Groupe 2

Gabriel Lebovic 2024-2025

# Contexte

## Présentation des données

## Regroupement de données

Pour la variable « marital-status » : le regroupement a été fait de la manière suivante :

* Marié 🡪 Married-AF-spouse ; Married-civ-spouse ; Married-spouse-absent
* Plus marié 🡪 Divorced ; Separated ; Widowed
* Jamais marié 🡪 Never-married

Pour la variable native-country, elle est représentée par la modalité « United-States » à hauteur de 90 % et ensuite aucunes autres modalités dépassent les 2 %, donc on décide de regrouper tous les pays ensemble et de composer la modalité « Le reste du monde », ce qui nous fait une modalité qui contient maintenant 10 % de notre population, ce qui est acceptable.

# Contexte

## Objectifs

Objectif clair pour décider des métriques à utiliser et leur seuil

## Description des données

## Traitement des données

# Analyse exploratoire

## Statistiques descriptives

(Rapide pour avoir une idée de la répartition de notre variable cible)

## Corrélations

(Important pour le choix de nos variables dans le modèle)

Matrice de corrélation

## Sélection des variables

Justification du choix ou non choix des variables

# Modèles et performances

## Premier cycle

### Entrainement des modèles

Choix des paramètre et présentation des modèles

### Métriques

Justification du choix des métriques en fonction des objectifs énoncés avant et calculs

### Interprétation

Interprétation des métriques obtenus et justifier les axes d'améliorations pour les prochains modèles

## Deuxième cycle

### Modifications apportées

Changement de variables, de données, de paramètres…

### Métriques

Calculs et interprétations des métriques

### Comparaisons

Comparaison entre modèles du premier cycle et entre reg log et arbre

## A l’infini en fonction du nb de cylce

# Modèle retenu

## Présentation modèles finales

Présentation et justification de pourquoi on retient ces modèles

## Prédictions du revenu

Prédiction des classes de revenus sur la nouvelle data

# Conclusion

Comparaison entre les deux modèles (arbre et reglog)

Force et faiblesse de chaque modèle