

PRÉDICTION DES CHARGES D'ASSURANCE : MODÈLE

Fatima, Umberto, Flora

Objectif

- **Contexte** : élaboration d'un modèle pour prédire les charges d'assurance de nouveaux clients
- **Ce qui a été fait** : analyse de données de clients existants
- **Observations principales** :
 - Fort impact du tabagisme sur les charges
 - Impact de l'obésité chez les fumeurs
 - Impact linéaire de l'âge
- **Objectif actuel** : produire un modèle de régression linéaire performant

Pipeline pour la modélisation

**Préparation
des données**



Aucune valeur
manquante

1 doublon
supprimé

**Encodage des
variables**



Label encoding
pour sex et
smoker (0, 1)

One-Hot encoding
sur region

**Division des
données**



Entraînement : 80 %
Test : 20 %

stratify sur smoker

**Test et
évaluation
du modèle**

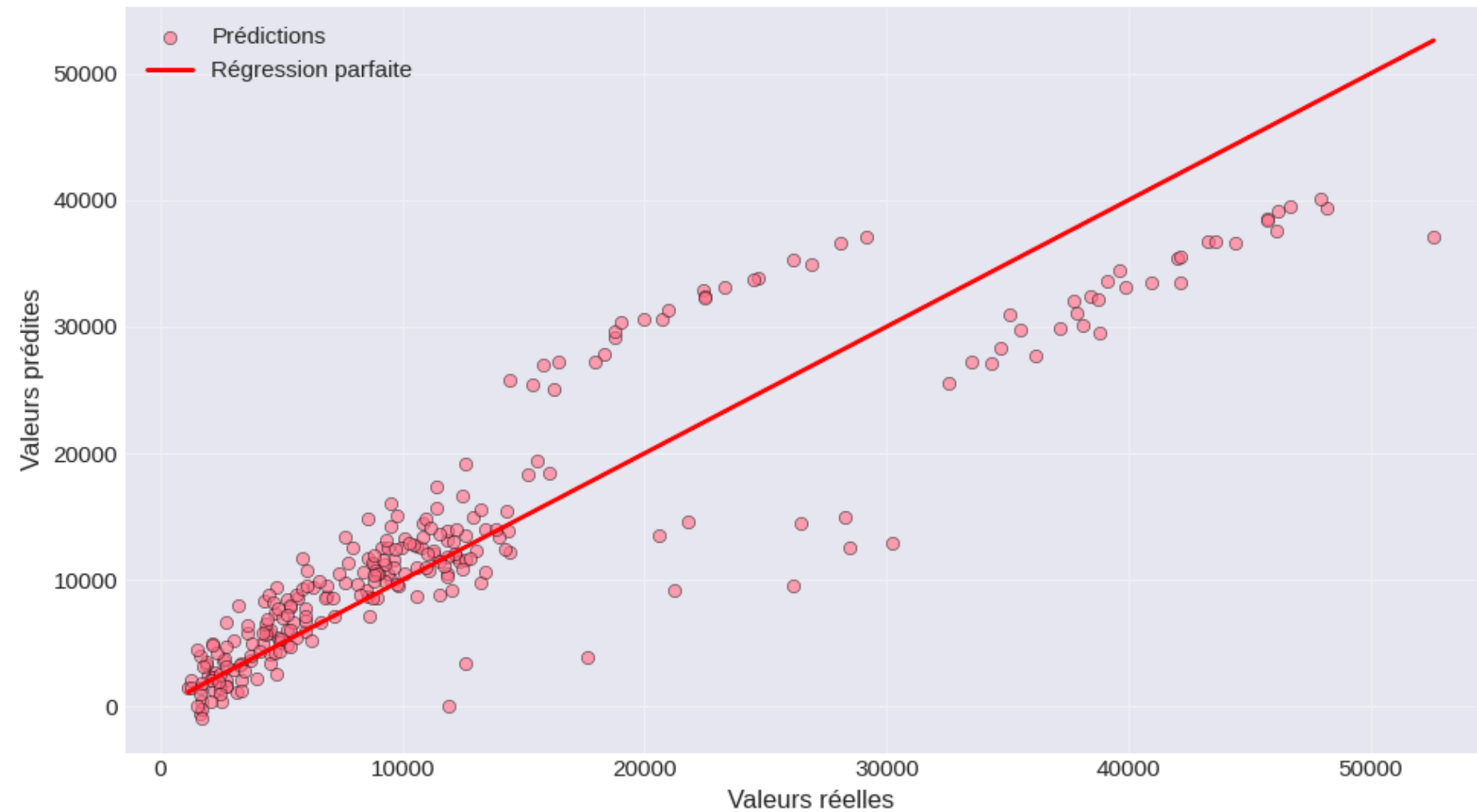


Mesure de la
performance avec
R², MAE, RMSE

Visualisation des
prédictions

Résultats du premier modèle

Régression linéaire sans log(charges)
(score R2 : 0.8205)

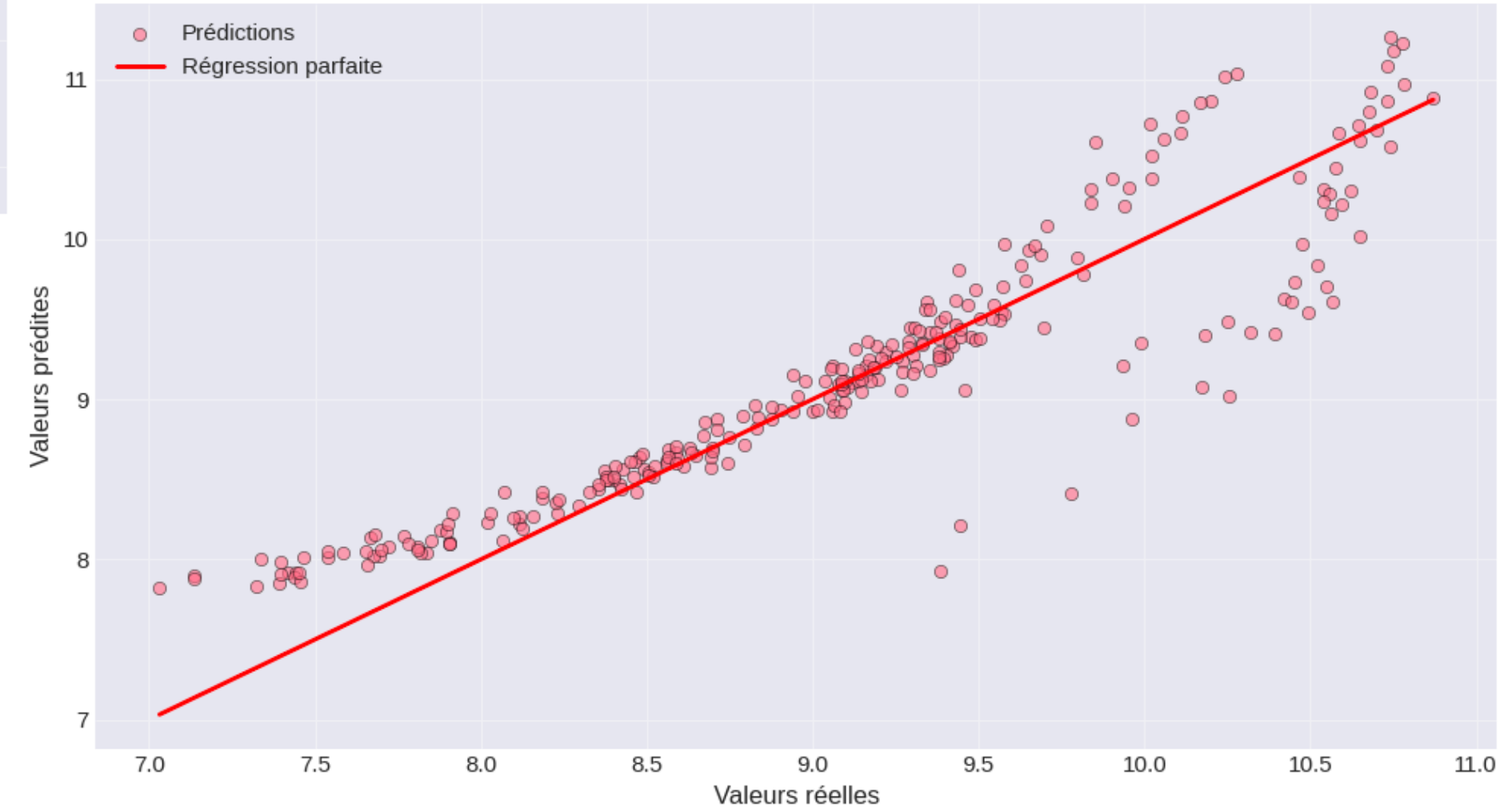


Mesure des performances :

82 % de précision

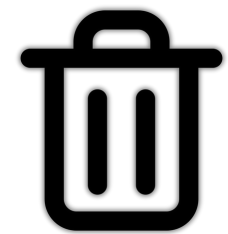
(MAE = 3598 \$ | RMSE = 5087 \$)

Régression linéaire avec log(charges)
(score R2 : 0.8376)



Performance avec log(charges) :
84 % de précision

Améliorations : choix techniques



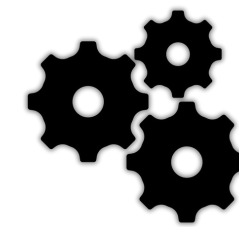
Suppression des régions



Ajout d'interactions :
smoker * bmi
smoker * is_obese
Suppression du bmi



Comparaison avec et
sans log sur les charges
→ Pas de log(charges)



Standardisation de
age, children
après le split pour
éviter la fuite de
données

Résultats après amélioration

Ancien score :
R2 = 84 %

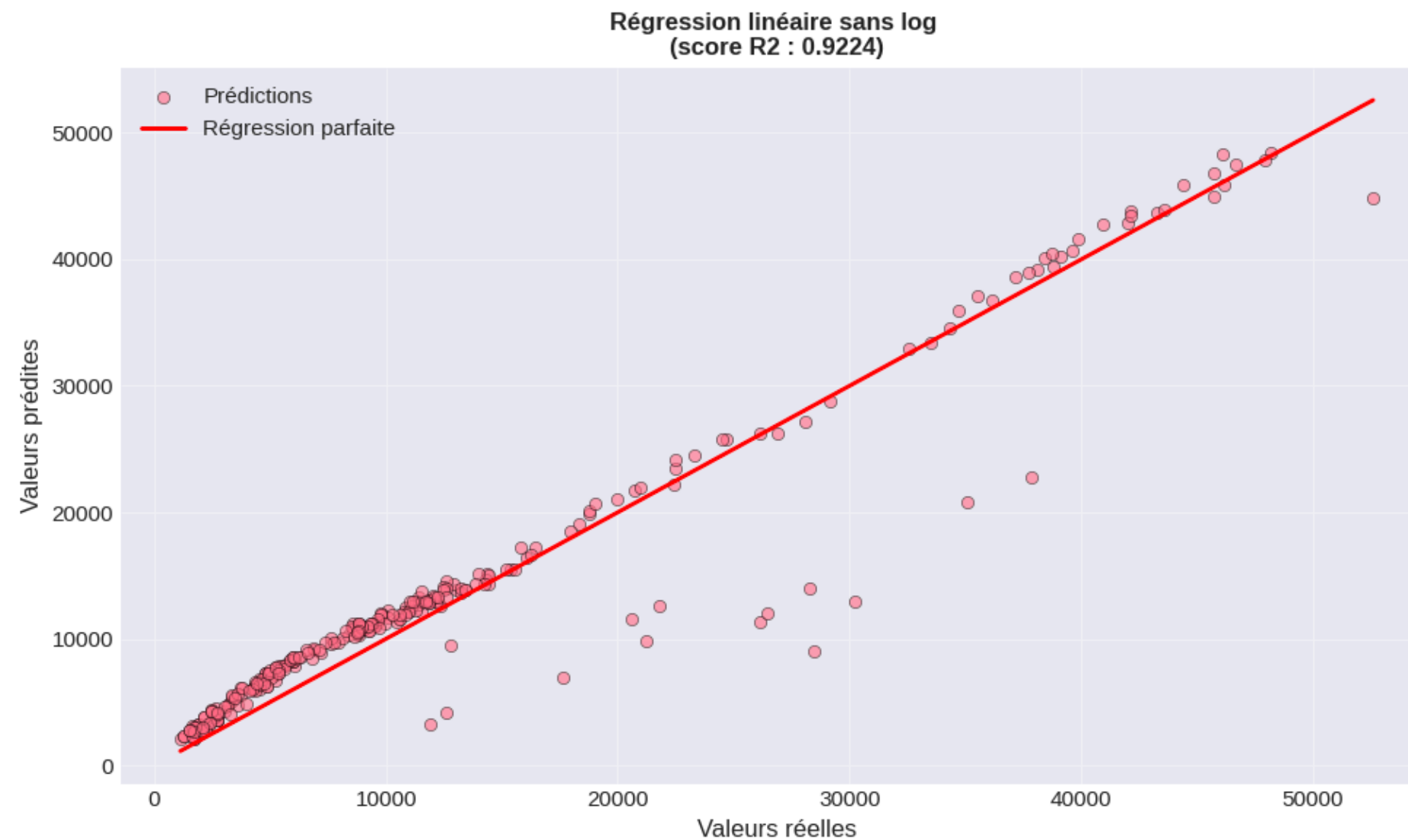


Modèle amélioré :

R2 = 92%

MAE = 2027 \$

RMSE = 3345 \$



Performance du modèle

Analyse
des résidus

Ridge
& Lasso

Polynomial
features

Cross-
validation

Automatisation et déploiement

Pipeline
automatisée

Application
Streamlit

**Axes
d'amélioration**

Merci pour votre attention !

Des questions ?