

SOULEYMANE SALL

Data Scientist

Paris, Île-de-France 07 53 44 95 60 sallsouleymane2207@gmail.com

Portfolio Medium GitHub LinkedIn

PROFIL

Data Scientist junior titulaire d'un Master en Statistique et Économétrie Appliquée. Maîtrise les étapes clés du cycle de vie des modèles de Machine Learning, du feature engineering au déploiement, avec une base solide en modélisation statistique et prédictive. Première expérience du déploiement via API, capable de passer du prototype à la production. Forte appétence pour l'intégration des pratiques CI/CD et MLOps.

EXPÉRIENCE

Data Scientist Micropole (groupe Talan)

Mars 2025 – Septembre 2025 / Paris

- Conception et déploiement d'un pipeline complet de segmentation client sur GCP (Vertex AI, BigQuery).
- Construction d'une base de données client enrichie à partir de tables brutes, permettant d'analyser la récence, la fréquence et la valeur des achats.
- Réalisation d'une analyse d'association produits (Market Basket Analysis) pour identifier des opportunités de cross-sell et d'upsell pertinentes.
- Production des rapports et visualisations pour les équipes marketing.

COMPÉTENCES

- Programmation** : Python (NumPy, Pandas, Scikit-learn), R, SAS, SQL
- Machine Learning** : Scoring, modélisation prédictive, régression, arbres de décision
- Outils** : GCP (Vertex AI, BigQuery), Dataiku, Power BI, Excel
- MLOps** : CI/CD, MLflow, Docker, API (FastAPI, Flask)

PROJETS

Scoring crédit

Mobilize financial services

- Modélisation d'un score d'octroi de crédit par régression logistique.
- Création d'une grille de score pour évaluer les demandes de crédit.
- Challenger la régression logistique par le SVM et le XGBoost.

Modèle de détection de phishing

Mailinblack

- Extraction et création de prédicteurs à partir des URLs, incluant l'utilisation de TF-IDF et bag-of-words.
- Optimisation du modèle via validation croisée, atteignant 95% de précision sur l'échantillon de test.

Détection de fraude

Projet académique

- Construction d'un modèle de détection des transactions frauduleuses à l'aide de la régression logistique et Random forest.
- Utilisation de techniques de rééquilibrage des données (SMOTE, ADASYN) pour mieux représenter les cas de fraude, ce qui a permis d'améliorer la détection de 25% et de réduire les erreurs de 20%.

FORMATION

Master Économétrie et Statistique Appliquée Université d'Orléans

2023–2025

Licence d'Économie Université de Bourgogne

2019–2022

CERTIFICATIONS

SAS Certified Statistical Business Analyst | SAS Base Programming | Python for Data Science | Dataiku (core Designer & ML Practitioner)

LANGUES

Français (natif) | Anglais (TOEIC 895 – C1) | Espagnol (notions)

INTÉRÊTS

Lecture, cuisine, sport