RIMA HAJOU

Senior Data Scientist

📍 Paris, France ✉ rimahajou@gmail.com
☎ +33 06 49 34 74 39 🌐 French citizen

SUMMARY

Experienced senior data scientist with over six years of expertise in implementing machine learning projects from inception to production. Proficient in addressing diverse challenges, including time series (forecasting and clustering), computer vision (image classification and object detection), and NLP (text entity recognition and text classification). My extensive background extends across various industries, including energy, oil and gas, maritime shipping, transportation, commodities, manufacturing, finance, and insurance.

SKILLS

Development	Python, PySpark, Pandas, Matplotlib, Scikit-Learn, HuggingFace, PyTorch, SQL, Object Oriented Programming
MLOps:	Experiment Tracking, Model Serving & Monitoring, MLflow, Tensorboard, Git, Code Reviews, Unit Testing, CI/CD Pipelines, Hydra, Docker
Cloud:	Azure, Databricks, AWS
Languages:	English, French & Arabic

EDUCATION

2017 - 2018	**M.Sc. in Data Science** Paris Saclay University, France	**Paris, France**
2011 - 2016	**B.Sc. in Computer and Telecommunication Engineering** Lebanese University, Faculty of engineering	**Beirut, Lebanon**

EXPERIENCE

11/2021 – Present **Senior Data Scientist** **Quantmetry, Part of Capgemini Invent - Paris, France**

Document Information Extraction using LayoutLM Transformers — **Insurance Sector**

- Led the development of an end-to-end pipeline for reading, analyzing, and extracting personal information from unstructured documents, including PDFs and images.
- Conducted a comprehensive benchmark and comparison of OCR techniques, including Azure Read API, Textract, Tesseract, and EasyOCR.
- Implemented code optimization strategies using Python and PySpark to enhance efficiency.
- Successfully implemented Key Information Extraction models leveraging state-of-the-art Transformers, specifically LayoutLM from Huggingface.
- Led the implementation, fine-tuning, and deployment to production of image classification algorithms (MobileNet, ResNet, and EfficientNet) to filter documents and reduce operational costs.
- Managed and developed packages from Proof of Concept to production, ensuring seamless integration.
- Established packaging and set up Continuous Integration / Continuous Deployment (CI/CD) with integrated unit tests, black, flake8, and mypy to maintain code quality.
- Developed a monitoring package utilizing MLFlow to track and manage the lifecycle of ML models.
- Coordinated closely with the business team to implement continuous improvements based on feedback and evolving requirements from clients.

`Python, PySpark, HuggingFace, Transformers, LayoutLM, Entity Recognition, Computer Vision, Azure, Databricks, CI/CD, MLFlow`

Passenger flow forecasting for Paris airports using ensemble ML models — **Services Sector**

- Conducted comprehensive analysis and processing of extensive airport data sets, systematically identifying trends critical for effective modeling.
- Led the identification and implementation of optimal approaches to forecast traffic flow, contributing to enhanced predictive accuracy.
- Demonstrated expertise in evaluating and seamlessly integrating diverse exogenous data types, thereby significantly improving overall model performance.
- Proactively managed and organized client meetings, fostering effective communication channels to collect valuable feedback. Utilized insights gathered to iteratively enhance model accuracy aligned with specific business needs.
- Successfully implemented advanced machine learning forecasting algorithms, leveraging the power of XGBoost, to deliver precise and reliable short- and medium-term traffic forecasts. This involved a strategic combination of algorithmic sophistication and domain-specific insights.

`Python, Time Series Forecasting, Hydra, MLFlow, LightGBM, Git`

Benchmark Deep learning for Time Series Forecasting — R&D

- Orchestrated and facilitated learning sessions to delve into deep learning research papers focused on advancing time series forecasting methodologies.
- Lead and manage of a 3-person data science team in the development of a robust benchmark, ensuring a collaborative and results-driven approach.
- Conducted thorough code reviews, merging changes, and consistently maintained the project's codebase to uphold high standards of code quality and project integrity.
- Implemented state-of-the-art Informer and TFT deep learning models, introducing new layers to incorporate exogenous dynamic and static variables, thereby expanding the benchmark's capabilities.
- Developed and executed time series temporal training validation protocols for deep learning models, ensuring the reliability and accuracy of the benchmarked results.
- Systematically compared benchmark results to LGBM, analyzing performance gains, training/inference times, and GPU costs, providing valuable insights for optimizing forecasting models.

`Python, Time Series Forecasting, Hydra, MLFlow, XGboost, Git, GluonTS, PyTorch`

Detection and classification of policies in insurance contracts using BERT — Insurance Sector

- Loaded and processed insurance contracts from PDFs using the Textract AWS service, ensuring efficient extraction of relevant data.
- Implemented pre-processing techniques, leveraging OCR metadata, to detect clauses within insurance contracts, enhancing the accuracy of subsequent analyses.
- Developed and fine-tuned classification models to predict policies at the line, paragraph, and page levels, optimizing granularity in policy identification.
- Orchestrated the development and containerization of an end-to-end Streamlit application using Docker. This empowered business users to seamlessly test and utilize the tool on a daily basis, fostering practicality and accessibility.

`Python, OCR, Textract, AWS, BERT, NLP, Streamlit, Topic Modelling`

Traffic Forecasting for Public Transportation — Transportation Sector

- Implemented a variety of techniques to improve time series imputation for missing data, enhancing the quality of data used for forecasting.
- Identified, processed, and analyzed exogenous features from diverse sources (weather, services) to augment the predictive capabilities of the forecasting models.
- Implemented and benchmarked machine learning models, focusing on short-term traffic forecasts at the train/station and time level, ensuring precision in predictions.
- Orchestrated workshops with clients to present forecasting results, facilitating a collaborative exchange of insights and feedback for continuous improvement.

`Python, Time Series Forecasting, Time Series Imputation, Kedro, AWS, Pandas, Matplotlib, Sklearn`

11/2018 – 10/2021 **Data Scientist** — **Quantcube Technology - Paris, France**

Nowcasting trade Economic indicators — International Trade Sector

- Analyzed shipping location data to identify changes in macroeconomic behavior, creating insightful macro-economic indicators for trade sectors. These indicators serve as valuable signals for informed country-based investment decisions.
- Initiated and led an end-to-end project, overseeing the migration to new types of shipping data. Conducted comprehensive benchmarking of different providers' data based on quality, coverage, latency, and data completeness to ensure data accuracy and reliability.
- Implemented, processed, and optimized data extraction and feature engineering from high-frequency AIS ships geospatial data. Leveraged tools such as PySpark, Pandas, and GeoPandas to handle the complexities of the data efficiently.

`Python, PySpark, Time Series analysis, AWS, Databricks`

Nowcasting commodities exports and imports — Commodities Sector

- Conducted market research to identify optimal techniques for exploring and leveraging maritime data, ensuring a data-driven approach in the commodities sector.
- Utilized spatial data and maritime metadata to create indicators for nowcasting imports and exports of crucial commodities such as crude oil, gas, coal, iron ore, and agricultural products. This facilitated timely insights for top exporters and importers.
- Developed and backtested maritime datasets, generating indicators that effectively represent changes in commodities exports, including but not limited to oil, coal, iron, and agricultural products. This process contributed to robust and reliable nowcasting models.

`PySpark, AWS, Time Series Analysis, Clustering, Spatial Data, Maritime Shipping Data`

Prediction of heavy fuel sales at Bunker ports — Energy Sector

- Conducted comprehensive research and market analysis on main bunker fuel ports, evaluating their impact on the heavy fuel market. This informed strategic decision-making in the energy sector.
- Analyzed spatial trajectories of ships, extracting meaningful patterns to enhance the understanding of maritime activities and their relation to heavy fuel sales.
- Estimated bunker fuel sales at one of the world's largest bunkering ports by leveraging ships' AIS data and trajectory patterns. This involved a meticulous examination of data to ensure accurate predictions.
- Performed data analysis on localization data using geopandas and folium libraries, providing valuable insights into the geographical aspects of heavy fuel sales at bunker ports.

`Spatial Data, Oil and gas, Spatial Data, PySpark, Folium`

Long term urban growth forecasting using Satellite Data — Urban Development Sector

- Implemented YOLO image segmentation techniques for urban feature extraction from satellite images.
- Analyzed and prepared Sentinel-2 imagery for high-quality data, enhancing segmentation accuracy.
- Extracted labeled data from OpenStreetMap for a comprehensive dataset, thus improving model training.
- Estimated urban growth in major cities by analyzing the evolution of urban segments, forecasting long-term development trends.

`Computer Vision, Satellite Images, Image Segmentation, YOLO, openstreetmap, Deep Learning`

Vehicle traffic forecasting for European highways using toll data — Transportation Sector

- Processed toll data, integrating highway metadata for nuanced insights into traffic patterns and contributing to a detailed understanding of the data.
- Applied advanced time series clustering to group highways based on toll data, allowing a nuanced forecasting approach considering distinct segment attributes.
- Implemented ensemble models (LightGBM, Random Forest) for accurate medium- and long-term traffic forecasts. This multi-model strategy enhanced prediction reliability by capturing diverse toll data patterns.
- Collaborated on presenting the project as a poster at the Banque de France and Bocconi University conference on Alternative Datasets, showcasing innovation in macroeconomic analysis with alternative datasets and receiving valuable industry feedback.

`Time Series Forecasting, Clustering, Spatial Data, PySpark, AWS`

04/2018– 09/2018 **Data Scientist Intern** — **Société Générale - Paris, France**

- Modeled default probabilities for personal and SME accounts using R (Logistic Regression, Ensembles).
- Analyzed model stability over time, ensuring reliability.
- Developed visualizations with RShiny for effective communication.
- Optimized in-production models through feature selection.
- Analyzed stability impact across various risk classes for different account types.

09/2016– 09/2017 **Data Consultant** — **Paragon Shift - Beirut, Lebanon**

- Manipulated, cleaned, and processed data to develop insightful analysis dashboards.
- Implemented technical solutions using Qlik Sense and Power BI, aligning with functional requirements.
- Conducted thorough raw data analysis, providing clients with actionable conclusions & recommendations.

`Data Vizualisation, PowerBI, Qliksense`

COMMUNITY PROJECTS

2023 **M6 Hackathon: 3rd prize in Forecasting of Financial Assets Returns** — **Link to blog post | Link to details**

- Successfully competed in a year-long hackathon with 12 submissions facing more than 200 participant, showcasing perseverance, dedication, and consistent efforts throughout the competition.
- Managed the project as a side project, effectively balancing responsibilities, and incorporated it into the R&D program, showcasing the project's relevance and contribution to ongoing innovation efforts.

`Time Series Forecasting, Financial data, Investment strategies, Stock market`

2022 - Present **Co-organizer of Paris Data Ladies Meetup** — **Link to meetup profile**

- Coordinate meetups aimed at empowering women in AI to showcase their work (Data Scientists, Data Engineers, Product Owners)
- Reach out to and arrange venues, ensuring smooth logistics and event preparation