Tommaso Baroni

+39 340 2111563 | tommaso1.baroni@gmail.com | linkedin.com/in/tommasobaroni | github.com/tommasobbb

Experience

Data Science Intern Nov. 2024 – May 2025
Deix Srl, Milan

- Co-led development of a Transformer-based Credit Risk Model (AUC 0.89), building pipelines to process data from over 300,000 companies. The model was trained using a self-supervised approach, similar to BERT, and we used attention-score analysis for interpretability to explain model predictions
- Implemented a multi-agent RAG system with LangGraph and developed the FastAPI/SQLAlchemy/PostgreSQL credit-management API
- Extended an internal tsfresh-based Python library to automate time-series quality checks on InfluxDB data, adding anomaly detection and seasonal-pattern analysis

Master Thesis Student Feb. 2024 – Oct. 2024
Ericsson, Stockholm

- Developed a Transformer-based time-series model for power-amplifier signals, designing custom embeddings to capture historical patterns and memory effects
- Surpassed RNN benchmarks by 2.3%, matching performance of Ericsson's proprietary solutions

Education

EIT Digital Double MSc, Data Science Sep. 2022 – Apr. 2025
Politecnico di Milano & U. Politécnica de Madrid **GPA:** 105/110

- **Selected Coursework:** Machine Learning, Deep Learning, Data Mining & Time Series, Complex Data in Health, Algorithms & Data Structures

BSc, Automation Engineering Sep. 2019 – Sep. 2022
Politecnico di Milano **GPA:** 105/110

- **Selected Coursework:** Probability and Statistics, Operations Research, Information Systems

Projects

Expected Goals (xG) Model from StatsBomb Open Data Jul. 2025 – Sep. 2025

- Developed an end-to-end machine learning pipeline to predict expected goals (xG), performing data extraction, feature engineering (including geometric and freeze-frame features), and training various models (Logistic Regression, Random Forest, XGBoost and an MLP) with cross-validation and grid search
- Calibrated the trained models using isotonic regression, and evaluated their performance using metrics such as Brier Score and ROC-AUC. Achieved performance comparable to StatsBomb's proprietary model, as demonstrated by the calibration curve

Spark-Optimized Parallel ML Algorithms Sep. 2023 – Jan. 2024

- Implemented parallel Logistic Regression and K-Means pipelines in PySpark, using only low-level Spark RDD transformations for data ingestion, normalization, gradient-descent training, accuracy computation, and cross-validation
- Benchmarked scalability (1–4 cores) on botnet traffic classification and MNIST clustering, achieving near-linear speedup and visualizing performance, convergence and evaluation-metric curves

Technical Skills

Languages: Python, SQL, Java, C, MATLAB
Frameworks & Libraries: PyTorch, scikit-learn, PySpark, LangChain, SQLAlchemy, FastAPI, pandas, NumPy
Databases: PostgreSQL, MySQL, InfluxDB
Tools & DevOps: Git, Docker, Linux, GitHub Actions, VS Code