

Giuseppe Muschetta

Machine Learning Engineer · Data Scientist · Computer Scientist

Based in Pisa, Italy · open to Machine Learning Engineer and Data Scientist roles, in Italy or abroad

PROFILE

Machine Learning Engineer and Data Scientist completing an MSc in Computer Science (Data Science & Business Informatics) at the University of Pisa. I build end-to-end models on real, large-scale data: a custom interior-point solver up to 264.9× faster than standard tools, explainable models with SHAP, NLP from TF-IDF to transformers, and PySpark pipelines on millions of records. Before computer science I spent 10+ years in sales, where I learned to turn complex results into clear decisions for non-technical people.

CORE STACK Python · PyTorch · TensorFlow · scikit-learn · XGBoost · Hugging Face · SHAP · PySpark · SQL · Power BI · Docker

SELECTED PROJECTS · ALL OPEN SOURCE ON GITHUB

High-Performance Mehrotra Lasso Solver ↗

Custom Mehrotra predictor-corrector interior-point solver for constrained Lasso, from the KKT conditions to a structure-exploiting implementation. **Up to 264.9× faster** than OSQP/SCS via CVXPY, with near-identical solutions, up to 2,000 features. Benchmarked on four datasets, including 148,432 IMDb samples.

Python · NumPy · SciPy · CVXPY · convex optimization

SoundPulse — Explainable ML on Spotify ↗

End-to-end pipeline on 114,000 tracks: anomaly detection, clustering validated by a 4.6M-parameter deep network (97.5% agreement), ensembles with **18× the R²** of linear baselines and 0.90 macro ROC AUC on 10 genres. SMOTE raised recall on the rare explicit class by 59%.

TensorFlow · Keras · XGBoost · imbalanced-learn · SHAP

Planet-Scale Earthquake Mining with PySpark ↗

3.27 million seismic records (1970–2019) processed in a distributed way: regression-based imputation instead of dropping data, K-Means clusters that isolate deep events down to 735 km, Random Forest magnitude classification and interactive Folium maps.

PySpark · Spark MLlib · Random Forest · Folium

IMDb: Advanced Data Mining, Explainable AI & Time Series ↗

My largest project: **149,531 records** and 1,134 box-office time series. Consensus anomaly detection (1,089 anomalies), learning under a 96:4 imbalance, ensembles and neural networks explained with SHAP, DTW clustering and MiniROCKET.

XGBoost · SHAP · tslearn · sktime · STUMPY

NLP: Amazon Sentiment, from Machine Learning to Transformers ↗

568,454 reviews, four generations of NLP: Naive Bayes, linear SVM, a custom 2.66M-parameter CNN-BiLSTM on Word2Vec and DistilBERT, with **ROC AUC up to 0.98** and detailed error analysis.

Keras · PyTorch · Hugging Face · Gensim · spaCy

Native Categorical Multiway-Split Decision Trees ↗

BSc thesis: an original decision tree written from scratch, with native categorical attributes and multiway splits, **benchmarked on 10 datasets** against scikit-learn (51,840 evaluations).

Python · NumPy · pandas · algorithm design

EXPERIENCE

Machine Learning Engineer

2023 – present

University of Pisa · applied data science and machine learning projects, alongside the MSc

- Six end-to-end projects in optimization, explainable machine learning, NLP and big data, on datasets of up to 3.27 million records; all open source.
- Business intelligence: data warehousing, ETL with SSIS, OLAP cubes with SSAS, MDX queries and Power BI dashboards.
- Full workflow: data cleaning and feature engineering, model selection, evaluation beyond accuracy, explainability, technical reports of up to 32 pages.

Independent Software Development

2021 – 2023

Alongside the BSc · systems and networking software, all open source

- WORTH, a distributed Kanban platform in Java: multithreaded TCP server, RMI callbacks, UDP multicast chat, persistent storage.
- A multi-process supermarket simulator in pure C (POSIX threads, sockets, signals), NIO servers and an interpreter in OCaml.

Sales Representative

10+ years, until 2018

Telecommunications, energy & retail · Bari area

- Outbound telesales for Vodafone, Eni, Fastweb, 3 and Wind (FlipCall on behalf of Comdata; E-Care), from first contact to signed contract.
- Retail at GameStop in several Bari-area shopping centers, including store openings at Auchan Casamassima; also a computer store and Foot Locker.
- Full sales cycle against daily and monthly targets: needs analysis, objection handling, negotiation and closing; explained technical products in plain words.

EDUCATION

MSc in Computer Science — Data Science & Business Informatics

2023 - in progress

University of Pisa

- Exams in the study plan: Machine Learning, Data Mining, Distributed Data Analysis and Mining, Natural Language Processing, Statistics for Data Science, Optimization for Data Science, Decision Support Systems, Business Process Models.

BSc in Computer Science

2018 - 2023

University of Pisa · thesis: an original multiway-split decision tree with native categorical support, written from scratch in Python

- Exams in the study plan include: Algorithms, Programming I & II, Operating Systems, Computer Networks, Databases, Software Engineering, Artificial Intelligence, Cryptography, Computability and Complexity, Information Theory, Calculus, Linear Algebra, Probability and Statistics, Numerical Methods, Operations Research.

Scientific High School Diploma

E. Amaldi Scientific High School, Bitetto (Bari)

MORE PROJECTS

IMDb Data Mining & Predictive Analytics — clustering, regression ($R^2 = 0.75$) and FP-Growth rules powering an interpretable classifier.

BPMN & Petri Nets — a business process modeled in BPMN and formally verified with Petri nets.

Yeo-Johnson Likelihood Inference in R — maximum likelihood estimation, Hessian-based uncertainty and likelihood-ratio tests.

github.com/giuseppemuschetta — 24 public repositories, from concurrency in C to machine learning.

SKILLS

MACHINE LEARNING & AI	PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, Hugging Face, SHAP, imbalanced-learn, spaCy, NLTK, Gensim
BUSINESS INTELLIGENCE	Data Warehousing, ETL, SQL Server, SSIS, SSAS, OLAP Cubes, MDX, Power BI, Decision Support Systems
LANGUAGES	Python, R, SQL, Java, C, C++, OCaml
DATA & VISUALIZATION	Pandas, Polars, NumPy, SciPy, CVXPY, Matplotlib, Seaborn, Folium, Jupyter
BIG DATA & DATABASES	Apache Spark, PySpark, Spark MLlib, MySQL, SQL Server
SYSTEMS & TOOLS	Linux, Multithreading, Sockets, Git, Docker
MATHEMATICS	Calculus, Linear Algebra, Probability, Statistics, Numerical Methods, Convex Optimization, Operations Research, Graph Theory
COMMUNICATION & SALES	Communication, Negotiation, Objection Handling, Persuasion, Closing Deals, Problem Solving

LANGUAGES

ITALIAN	Native
ENGLISH	Full professional proficiency · B2 certificate, CLI, University of Pisa

AVAILABLE ON REQUEST

Phone number and degree certificate.

I authorize the processing of my personal data pursuant to Regulation (EU) 2016/679 (GDPR) and the applicable national legislation.