Tommaso **Guerrini**

Data Science Lead - AI expert - Kaggle Master

Milano, Italy

(+39) 338-6590515 | guerrinitom@gmail.com | tomwarrens | tommaso-guerrini

- **Data Scientist** with 9 years of experience across many different industries and companies.
- Highly experienced in tabular data problems, particularly binary classification tasks such as credit risk and anomaly detection.
- Experience deploying and maintaining machine learning models into production environments.
- Proficient in Python and PySpark, experienced in PyTorch.
- Experience with all major Cloud Providers, especially Google Cloud Platform and AWS.
- Experienced in working with peers and mentoring younger colleagues.
- Active member of Kaggle, StackOverflow, CrunchDAO, StackExchange, GitHub, and Leetcode communities.

Experience

Tabular Data Lead at CardoAI

Jul. 2022 - Present

Lead all initiatives related to tabular data modeling and machine learning productization, focusing on credit risk, model governance, and data quality. Acted as the product owner responsible for the release and operationalization of models. Oversaw an associate, managing end-to-end model deployment and maintenance. Main projects and tasks performed included:

- Developed ensemble Probability of Default (PD) models (LightGBM, XGBoost, Isotonic Regression) for the US consumer and Italian SME markets, both at origination and throughout the life of the loan.
- Designed and deployed a model explainability framework using SHAP values, DeepChecks, and custom feature importance metrics such as average gain per split and first usage depth.
- Built an anomaly detection framework applied to all client data, automatically identifying issues such as missing values, invalid linear relationships, and broken association rules. This system significantly reduced client onboarding time and improved the quality and reliability of data transformations.
- Implemented automated retraining and monitoring pipelines via Kubeflow, MLflow, and Airflow.
- Collaborated with Data Engineering and DevOps teams to operationalize models through KServe and ArgoCD, ensuring CI/CD compliance.
- Mentored junior data scientists and contributed to defining internal model development lifecycle standards to enhance reproducibility and governance.

Tech: Kubernetes, Kubeflow (Notebooks, Pipelines, Katib), Docker, Git, ArgoCD, Lens. AWS (S3, ECR, EC2). Databricks. Airflow. Python (MLflow, Numpy, Pandas, Sklearn, LightGBM, Optuna, XGBoost, Flaml, PyCaret, PyTorch, Seaborn, KServe, re, spacy, nltk, shap), PySpark.

Machine Learning Scientist/Data Scientist at Generali Italia

Jan. 2020 - Jun. 2022

Worked on multiple AI initiatives in fraud detection and process automation, managing projects from data ingestion to production deployment on Google Cloud Platform. Responsible for developing performant models, integrating them into CI/CD pipelines, and improving data management workflows across teams. Main use cases and tasks performed included:

- Developed and deployed text classification models to automatically categorize documents, released as APIs via Google Cloud Functions and App Engine.
- Implemented graph-based fraud detection features using Neo4J, leveraging client–agent–third-party relationships to improve model accuracy by roughly 15%.
- Designed and deployed a PySpark-based feature store orchestrated via Jenkins and Airflow, enabling reusable and versioned feature pipelines.
- Built and optimized fraud detection models for car accidents using LightGBM and Optuna for hyperparameter tuning, deploying through Jenkins and Airflow.
- Collaborated cross-functionally to design reusable ML templates and CI/CD standards adopted by multiple teams within the innovation department.

Tech: Google Cloud Platform (Vertex AI, Dataproc, Compute Engine, Cloud Run, Cloud Functions, App Engine, BigQuery, Cloud Storage, Logging, Error Reporting)
Python (Numpy, Pandas, Sklearn, LightGBM, Optuna, PyCaret, PyTorch, Seaborn, Flask, SpaCy, NLTK, SHAP), PySpark, Neo4J, Jenkins, Airflow, Docker.

Data Scientist at Data Reply IT

Aug. 2016 - December 2019

- **Multinational company in the Automotive sector**: helped developing ***short and long term forecasting models*** along with the DS company team. The aim was to optimize the supply chain costs in one case and provide a data-driven help to strategy in the other. Models: Hierarchical Time Series, leaf-wise and level-wise Extreme Gradient Boosting.
  Tech: R (forecast), Python (Pandas, XGboost, Pyspark). AWS (S3, EMR, Athena, Glue). Agile, rotating pairs. Git. Unix.

- **Multinational company in the Manufacturing sector**: led a team of 3 people in releasing multiple supply chain and logistics business analytics dashboards. The aim was to ***provide*** demand planners and plant owners with hourly ***KPI reporting tools and anomaly detection alerts***. Models: Isolation Forests and Statistical tests.
  Tech: Pyspark, Microsoft Azure (DataLake, Hive, SQLDatabase)

- **Multinational company in the Telco sector**: worked with the Big Data team, along with the marketing team on customer care and proactivity, ***developing a model to predict customer care calls***.
  Tech: Pyspark, Pyspark ML, Pyspark MLLib, Python. Agile. Git. Unix.

- **Italian company in the Media sector**: led a team of 2 people on a project whose aim was ***real time forecasting of the company tv channels audience***. Our model led to more than 10 % performance improvement overall as well as 25 % in spiky periods.
  Tech: Python (XGBoost, LightGBM, Shap, Keras).

- **Multinational company in the Manufacturing sector**: helped the company in migrating from on premise infrastructure to Google Cloud Platform.
  Tech: Google Cloud Platform (GC Storage, BigQuery, DataStudio, AppEngine, Cron)

- **Italian Insurance Company**: worked in presales and PoC. Our aim was to develop an algorithm providing an estimate of the cost of repair of a damaged car from the picture of the car.
  Tech: Python (Keras, Tensorflow, Numpy, Pil, OpenCV)

- **Italian Retail Company**: built a ***sales forecasting model*** to forecast daily beverage supermarket sales for an italian retailer company.

Certifications and Awards

- **KAGGLE NOTEBOOK MASTER** (July 2021 – Present): Ranked as high as #163 worldwide and #1 in Italy in the code rankings on Kaggle. https://www.kaggle.com/tomwarrens/code
- **Winner – CLADAG 2019 Data Science Competition** (October 2019): Forecasting challenge jointly organized by Società Italiana di Statistica and TIM, won with a team of 2 fellow data scientists.
- **AWS Certified Cloud Practitioner** (July 2019).
- **NVIDIA: Fundamentals of Deep Learning for Multiple Data Types** (February 2019).
- **The Data Incubator Graduate Program** (June 2017).
- Active member of the data science community on Kaggle, StackOverflow, CrunchDAO, StackExchange, GitHub, and Leetcode.

Talks and Presentations

- **eXtreme Programming User Group Bergamo – From Chaos to Order: The Power of GitOps in Software and ML Development**
  May 2023: Joint talk on applying GitOps principles to streamline machine learning deployment and model lifecycle management.
  https://www.meetup.com/it-IT/xpugbg/events/293199666/
- **Data Engineering Milano – GitOps in AI: ArgoCD & Kubeflow Assemble**
  January 2023: Presentation on leveraging ArgoCD and Kubeflow for CI/CD in production ML pipelines.
  https://www.meetup.com/it-IT/data-engineering-italy/events/290673054/

Tools, Technologies, Courses

- **Programming Tools and Frameworks**: Python (Pandas, Numpy, Scikit-learn, LightGBM, XGBoost, PyTorch, Keras, TensorFlow, PySpark, PySparkML, PySpark MLLib, CXOracle), R (Tidyverse, forecast).
- **Big Data & Cloud Platforms**: PySpark; AWS (S3, EMR, EC2, Glue, Athena); Google Cloud Platform (Dataproc, Compute Engine, AI Platform, Vertex AI, Kubernetes Engine, Container Registry, Cloud Run, Cloud Functions, App Engine, BigQuery, Cloud Storage, Logging, Error Reporting); Microsoft Azure (DataLake, Hive, SQLDatabase).
- **Other Tools**: Unix, Git, Agile (Scrum and Kanban).
- **Courses & Teaching**: Instructor for courses in Deep Learning for satellite image segmentation and in Deep Reinforcement Learning for Atari game environments (25-person classes).

Education

M.Sc. in Mathematical Engineering, Applied Statistics *110/110*

Politecnico di Milano *Oct. 2014 - Apr. 2017*

- Relevant courses: Applied Statistics, Machine Learning, Artificial Intelligence, Algorithms and Parallel computing, Bayesian Statistics
- Final work: ”Machine Learning Models for Forecasting of Daily Supermarket Sales”

B.Sc. in Physics Engineering *95/110*

Politecnico di Milano *Oct. 2011 - Sep. 2014*

- Relevant courses: Quantum Physics, Linear Algebra, Partial Differential Equations