

Lorena Melo Santos

Data Scientist & ML Engineer

Belo Horizonte, Brazil — available for remote work | lorenamelo.engr@gmail.com
[linkedin.com/in/lorenameloengr](https://www.linkedin.com/in/lorenameloengr) | github.com/lorenamelos | lorenamelos.github.io

PROFILE

Data Scientist with two and a half years building forecasting, NLP and LLM systems in production environments, plus the data pipelines and warehouse modelling underneath them. Experience delivering into enterprise cloud infrastructure at BMW Financial Services Canada and Johnson & Johnson LATAM, with published research in NLP. Background in metallurgical engineering: a discipline of specification, tolerance and failure analysis, which shapes how the work gets validated.

EXPERIENCE

Data Scientist — Zallpy Digital Oct 2024 — Aug 2026

Allocated to enterprise clients, then to an internal commercial intelligence product.

- **BMW Financial Services Canada.** Built an end-to-end pipeline extracting 25+ structured attributes from customer survey comments — 3,094 comments processed in ~25 minutes at 92% accuracy against human annotation, raising feedback coverage from 30% to 100%. Architecture: Apache Iceberg into AWS Glue with PySpark, calling Claude Sonnet through an internal LLM gateway with OAuth 2.0.
- Built a three-tier sales forecasting system across ~1,800 retailer-model combinations (LightGBM / RandomForest / ElasticNet by data availability), reaching R^2 0.808 at daily granularity. Detected and eliminated data leakage in early versions and documented a validation rule flagging R^2 above 0.95 as a leakage indicator for future maintainers.
- Orchestrated both pipelines as AWS Step Functions state machines, with a weekly input-table refresh and monthly model retraining promoted to the production table backing stakeholder dashboards.
- Consolidated commercial finance, retailer and VIN-level data from on-premise SQL Server into a five-table warehouse asset; resolved Cartesian joins and duplicate records and implemented indexing that reduced query runtime by ~60% on datasets exceeding 15 million rows.
- **Johnson & Johnson LATAM.** Led development of a warehouse capacity forecasting model (XGBoost) across three distribution centres, predicting bin occupancy at SKU level — MAE 1.79 bins, WMAPE 6.2% at storage-type level, outperforming a moving-average baseline. 25 engineered features over 37 months of history. Delivered SHAP explainability and a Streamlit what-if simulator for scenario planning.
- **Internal product.** Built the classification core of a commercial intelligence system reading sales call transcripts: structured JSON output with option-list validation, per-dimension justification citing the source, and a pseudonymization pipeline validated as zero-cost in classification accuracy.

Research Fellow — University of Brasília (UnB) 2025 — Apr 2026

- Built a multilingual NLP pipeline with PyTorch and Hugging Face Transformers for longitudinal sentiment analysis of public reviews of Brazil's national public health application.
- Published research on LLM-based sentiment analysis at EnANPAD 2025.

Data Analyst — Konica Minolta Healthcare do Brasil Sep — Oct 2024

- Built SQL extraction workflows from PostgreSQL with data integrity checks, and analytical tooling in Streamlit and Plotly for exploring device performance across daily, monthly and yearly time scales.
- Time series analysis for trend and anomaly identification in operational monitoring data.

Data Science Teacher Assistant — Le Wagon May — Sep 2024

- Supported students through a data science bootcamp across Python, statistics, machine learning, deep learning and deployment fundamentals, debugging code live and giving structured feedback on projects.

SELECTED PROJECTS

- **Distill** — production data pipeline discovering remote roles, scoring fit and generating tailored resumes. Multi-source ingestion across four ATS APIs, three-layer deduplication, a validated state machine with an explicit transition table, idempotency by construction, per-record error isolation, automated backlog recovery with spend caps, and cost and lineage tracking per API call. 3,032 jobs processed. A deterministic pipeline with an LLM at two controlled points.
- **Merchant Social Intelligence Agents** — multi-agent system in LangGraph for the CloudWalk technical challenge, with routing derived from exploratory analysis of real message data, full test coverage and containerised delivery.
- **Apiculture RAG** — retrieval-augmented generation system over apiculture domain content.

SKILLS

Languages & core: Python, SQL, pandas, NumPy

ML & AI: scikit-learn, XGBoost, LightGBM, RandomForest, ElasticNet, PyTorch, TensorFlow, Hugging Face Transformers, multilingual BERT, SHAP, time series forecasting, feature engineering, NLP, prompt engineering, LangChain, Ollama, Claude API, RAG

Data & cloud: AWS Glue, AWS Step Functions, Apache Iceberg, S3, Athena, CloudWatch, Terraform, AWS Secrets Manager, PySpark, Databricks, Delta Lake, PostgreSQL, Supabase, SQL Server, Parquet, QuickSight, Tableau

Engineering: Git, GitHub Actions, Docker, Docker Compose, Streamlit, Plotly, pytest, ETL, data pipelines, pipeline orchestration, scheduled model retraining, state machines, observability

EDUCATION

MBA in Big Data — Descomplica (in progress)

Natural Language Processing — UFMG, graduate-level extension course

Data Science Bootcamp — Le Wagon

Metallurgical Engineering — PUC Minas · Engineering exchange, Queensland University of Technology (CNPq scholarship)

PUBLICATIONS

EnANPAD 2025 — research on LLM-based sentiment analysis.