CHAITHANYA VEMURI

Berlin, Germany | +49 15510944761 | chaithanyavemuri28@gmail.com
Portfolio | GitHub | LinkedIn

PROFILE

MSc Data Science student focused on Python/SQL data pipelines, recommendation and forecasting, and reproducible ML services. Built independent systems from data validation to offline evaluation and API serving. Seeking junior ML or Data Engineering roles. Based in Berlin; start date by agreement.

TECHNICAL SKILLS

Data engineering: Python, SQL, pandas, PostgreSQL, DuckDB, Parquet, dbt; PySpark and Redpanda streaming.

Machine learning: NumPy, scikit-learn; recommendation, forecasting, chronological validation. **ML systems:** MLflow, FastAPI, Docker, GitHub Actions.

SELECTED ENGINEERING PROJECTS

REES46 V2 | Recommendation Platform — GitHub

Python | SQL | DuckDB | Parquet | FastAPI | PostgreSQL | MLflow

- Processed **411.7M public events / 15.6M dataset users** into typed Parquet, canonical Silver and monthly Gold facts with DuckDB. Replaced a failed global aggregation (~30.6 GiB spill) with partitions/session buckets and resumable outputs.
- Hybrid retrieval/ranker benchmark: **MRR@20 0.02443 to 0.03556** (+45.6% vs popularity) on **5,000 held-out users**. Bounded training used ~1.50M interactions; added MLflow/FastAPI serving with a popularity fallback.

FulfillAI | Fulfillment Data & ML System — GitHub

Python | PostgreSQL | dbt | MLflow | FastAPI | Redpanda | PySpark

- Built a **50K synthetic-order** relational dataset, PostgreSQL/dbt staging and marts, quality checks and prediction-time feature contracts.
- Frozen chronological demand test (**63,501 synthetic rows**, Jun-Jul 2026): **69.59% WAPE vs 88.24%** rolling-28 baseline. Reported 21.14% relative reduction, high residual error and worse RMSE.
- Connected MLflow/FastAPI to reproducible artifacts; verified Redpanda/PySpark watermarking, checkpoint restart/resume and a PostgreSQL streaming sink.

OTTO Demand Forecasting | MSc Dissertation Research — GitHub

Python | Time Series | Chronological Validation | Model Evaluation

- Compared **eight approaches** on 672 hourly observations using three chronological folds, five neural seeds and an untouched 96-hour holdout.
- Selected weekly seasonality over the tested neural models: **13.93% mean CV WAPE**; holdout **9.80% carts / 11.34% orders** for 24-hour forecasts.

TRAINING

IntrnForte — 02/2024 - 07/2024

Python Self-Learning Internship | Remote

- Completed structured learning in Python, OOP, data structures and debugging; applied pandas, NumPy and Matplotlib for practical data handling.

EDUCATION

Gisma University of Applied Sciences — 08/2025 - 11/2026 (expected)

MSc Data Science, AI & Digital Business — Berlin, Germany

Dissertation: OTTO E-Commerce Demand Forecasting.

Sri Venkateswara College of Engineering — 10/2021 - 05/2025

BTech Electronics & Communication Engineering — Tirupati, India

LANGUAGE SKILLS

English: Currently learning toward C1-level proficiency

German: Currently learning toward B2-level proficiency