Sholto Armstrong

Machine Learning Engineer

sholto@sjnarmstrong.com | sjnarmstrong.com | github.com/sjnarmstrong | linkedin.com/in/sjnarmstrong | Johannesburg, South Africa (UTC+2)

SUMMARY

Machine Learning Engineer with 7+ years building and deploying production ML, RAG/LLM and real-time data systems across banking, IoT and retail. Led a 10-person AI team, built streaming data platforms processing 1M+ records per minute, and shipped production RAG systems on Kubernetes. MEng in Computer Engineering (cum laude). Core stack: Python, PyTorch, FastAPI, Kafka, Flink, Docker, Kubernetes, AWS/GCP.

EXPERIENCE

Machine Learning Engineer @ Capitec Bank

09/2023 – Present · Sandton, South Africa

capitecbank.co.za — the leading bank in South Africa by number of clients and user satisfaction

- Built and open-sourced Decider (github.com/capitec/dsp-decision-engine), a rules engine productionised for business credit granting, retail credit granting and fraud detection across 3 teams, executing rules on 7M+ transactions per day; presented the project at PyCon.
- Established the Kafka and Apache Flink patterns used for real-time stream processing of transactional data.
- Developed a real-time person-detection (computer vision) proof of concept on branch CCTV cameras, analysing where customers spend time in selected branches to target reductions in counter waiting times.
- Redefined the model deployment process around Docker containers, cutting deployment debugging from hours of hands-on support to minutes of self-service, and built CI/CD pipelines that automatically integrate jobs with the bank's scheduling platform.
- Drive hands-on enablement of multiple teams across the bank, supporting adoption of ML tooling and best practices.

AI Engineer (Part-time Consultant) @ CohesionX

02/2024 – Present · Remote, South Africa

prod.vectormind.chat — an AI startup building customised RAG chatbot systems

- Develop a production RAG chatbot platform built on LLMs, embeddings and vector search, alongside my full-time role.
- Migrated the platform from a single EC2 instance to Kubernetes, improving maintainability and reliability.
- Rebuilt the RAG ingestion and training flow on a serverless architecture, improving retrieval accuracy, reducing hallucinations and eliminating always-on compute costs.
- Implemented downloadable source references for chatbot answers, improving verifiability and user trust.

Product Engineering Manager (Machine Learning) @ IoT.NxT (Vodacom/Vodafone)

03/2023 – 09/2023 · Centurion, South Africa

iotnxt.com — an industry-agnostic IoT platform with Gartner recognition

- Led and managed a team of 10 developers delivering AI and data analytics products for the IoT platform, owning planning, delivery and stakeholder communication.

Senior Machine Learning Engineer (Team Lead) @ IoT.NxT (Vodacom/Vodafone)

06/2022 – 03/2023 · Centurion, South Africa

- Expanded the data analytics product into a low-cost multi-tenant solution scalable to thousands of clients, directly enabling sales to 6 high-value clients.
- Set up agile processes and CI/CD pipelines, reducing time from requirement to deployment by 75% (months to weeks) while improving product quality.
- Served on the platform architecture design board and hosted monthly AI knowledge-sharing sessions across the team.

Machine Learning Engineer @ IoT.NxT (Vodacom/Vodafone)

06/2021 – 06/2022 · Centurion, South Africa

- Built a data analytics platform — lakehouse (Apache Iceberg), BI reporting and stream processing — storing and processing 1M+ records per minute in near real-time.
- Developed a no-code platform enabling clients to easily create real-time ETL pipelines.
- Secured the full stack, including OAuth2 integration, and remediated 352 CVEs.

Junior Machine Learning Engineer @ IoT.NxT (Vodacom/Vodafone)

04/2020 – 06/2021 · Centurion, South Africa

- Deployed real-time ML predicting energy usage of air-conditioning units in cellular base stations, verifying control logic that reduced energy consumption by 8%.
- Built models to select optimal locations for alternative energy installations and to detect anomalous base-station telemetry, surfacing underperforming equipment.
- Created APIs and custom UIs serving model outputs, turning predictions into actionable insights.

Data Scientist @ Cognizance (MAC Mobile)

02/2019 – 03/2020 · Pretoria, South Africa

cognizance-vision.com — a retail analytics platform delivering real-time, actionable insights across the retail chain

- Trained a computer-vision model detecting front-facing shelf products for planogram compliance, achieving 97% accuracy.
- Optimised the model for serverless deployment, reducing detection latency from 15 minutes to 5 seconds.
- Built web tools for efficient collection and labelling of voice and image data, cutting labelling time per sample by 90%.

Full-Stack Developer @ MathU (Instasense/Uvirco)

01/2018 – 01/2019 · Pretoria, South Africa

- Sole developer at an AI-driven math-education startup: built an offline-capable mobile app for remote areas with unstable connectivity, plus a rewards-based data-collection programme that gathered thousands of sample math problems.

EDUCATION

Master of Engineering (MEng), Computer Engineering

01/2020 – 12/2023 · University of Pretoria

- Graduated cum laude; part-time enrolment with a research focus on scene understanding (computer vision).

Bachelor of Engineering Honours (BEngHons), Computer Engineering

01/2018 – 12/2019 · University of Pretoria

- Graduated cum laude; part-time enrolment with a focus on artificial intelligence.

Bachelor of Engineering (BEng), Computer Engineering

01/2014 – 12/2017 · University of Pretoria

- Graduated cum laude with a bronze merit award; silver merit awards in 2nd and 3rd year; Golden Key member.

SKILLS

Machine Learning: PyTorch, TensorFlow, scikit-learn, Pandas, computer vision, anomaly detection, time-series forecasting, model deployment, MLOps
LLMs & GenAI: RAG pipelines, LLMs, embeddings, vector databases, hallucination reduction
Data Engineering: Apache Flink, Kafka, Spark, Beam, Apache Iceberg, real-time streaming, ETL pipelines, lakehouse architecture
Programming Languages: Python, SQL, TypeScript, Java, C++, Rust, Kotlin
Backend: FastAPI, Flask, REST APIs, OAuth2
Databases & Storage: PostgreSQL, Redis, MongoDB
Infrastructure & DevOps: Docker, Kubernetes, Helm, AWS, GCP, CI/CD, serverless architectures
Frontend: React, Angular
Tools & Practices: Git, GitHub, GitLab, Azure DevOps, Agile