Rohan Anand

anandro@bu.edu | rh-anand.vercel.app | linkedin.com/in/rohan-h-anand | github.com/rohan-anand21 | (603) 233-4670

Experience

Data Engineer — Applied ML
Mar 2025 – Present

Dataeconomy — *Charlotte, NC*

- Built and shipped an implicit-feedback **ALS recommender** for a data marketplace; improved **Precision@10** by ∼50% vs. a popularity baseline via **Optuna** hyperparameter tuning.
- Productionized the model as a **Dockerized FastAPI** service on **AWS ECS**; created a **Jenkins** pipeline that automates retraining, offline evaluation, and deployment; model + config artifacts versioned for reproducibility.
- Built and operated **Kafka → Snowpipe → Snowflake** near–real-time ingestion pipeline to support internal data products; added **data-quality validation** and reports in Snowflake/Snowpark.
- Rewrote validation in **Great Expectations** within a PySpark ETL, adding new checks/constraints and reducing runtime by ∼70% on large datasets.
- Implemented historical data management with **Snowflake-managed Iceberg** and **SCD 0-4** (SQL + dbt) to ensure feature/history consistency for auditability.

AI Engineer
Sep 2024 – Jan 2025

BU Spark! — *Boston, MA*

- Delivered a retrieval + **generation** workflow that a client reported reduced initial investigation time by ∼80%; tracked offline retrieval quality and iterated on features and prompts.
- Ingested and normalized officer records (2011–2024) from heterogeneous spreadsheets (∼14 files, ∼50k rows, ∼30 columns); built a repeatable pipeline for cleaning, deduplication, and entity resolution; added incremental upserts to process new data
- Implemented an embeddings pipeline with **OpenAI** and a **Chroma** vector index (∼1,000 vectors, dim=1,536); applied chunking (∼500 tokens) and metadata enrichment (officer_id, allegation type, year) to improve retrieval quality and maintain traceability.
- Shipped a **FastAPI** service exposing `/recommend`: Chroma → retrieval → OpenAI generation, orchestrated with **LangChain**; supports CSV uploads for new records and writes request/response audit logs for review.
- Improved offline retrieval by ∼40% via feature augmentation (e.g., prior-offense context) and targeted prompt templates.
- Built geospatial Python maps to guide reviews across ∼23 areas and ∼13 time windows; overlaid counts/rates with filters (year, allegation type) for client working sessions.

Data Services Intern
May 2024 – Aug 2024

Axis Technology, LLC — *Boston, MA*

- Built a scalable **synthetic-data generator** in Python (1k → 100k+ rows) for PII detection; improved offline performance on a held-out real dataset while maintaining stable runtimes.
- Developed a **JSON schema parser/labeler** to auto-tag semantic column types (email/phone/ID), reducing manual tagging by **75%**; authored tests to evaluate **OpenSearch** table-similarity relevance and documented metrics/cases.

Data Analyst Intern
Jun 2022 – Aug 2022

AS Insurance Agency — *Manchester, NH*

- Consolidated customer and policy data (**2.5k+ docs**) with Pandas; wrote Snowflake SQL to segment customers; built **10–15** Power BI/Tableau dashboards for renewal targeting.
- Streamlined renewal outreach; contributed to **95% client retention** during the period.

Projects

Analyzing Boston's 311 Service Requests
Sep 2023 – Dec 2023

- Built an **automated daily API ingestion** for Boston 311 (**2.7M+ records** across 12 years); normalized raw responses into analysis-ready tables and added basic deduping and schema validation to keep the dataset clean.
- Developed interactive analyses in Jupyter (ipywidgets) and a **Power BI** map layered with the **CDC Social Vulnerability Index (SVI)**; examined trends by neighborhood, request type, submission source, and resolution time.
- Produced and presented a **Power BI** report showing differences across income levels; documented the data structure and the **daily refresh schedule**.

Maternal Health & Infant Outcomes
Sep 2023 – Dec 2023

- Built and deployed an interactive **Quarto** site on U.S. natality microdata; harmonized multi-year files (**200+ variables** → analysis set) via a reproducible **tidyverse** pipeline (cleaning, derivations, documentation); versioned on GitHub and deployed to Vercel.
- Ran two-sample comparisons and linear regression: in this dataset, **smoking was associated with lower birthweight and slightly lower APGAR** (effect sizes + 95% CIs). Model fit: adj. $R^2 \approx 0.50$ (birthweight), ≈ 0.004 (APGAR).
- Built state-level choropleths and race-stratified time-series; observed **declines in smoking prevalence (2014–2021)** and a **negative correlation** between state smoking prevalence and mean gestation length.

Skills

ML: scikit-learn, PyTorch, Optuna | **NLP/RAG:** OpenAI API, LangChain, Chroma | **Data & SQL:** Pandas, NumPy, dbt, Snowflake | **MLOps & Data Eng:** FastAPI, Docker, AWS (ECS/ECR, S3), Jenkins, Airflow, Kafka, PySpark, Great Expectations | **Storage/Search/BI:** PostgreSQL, OpenSearch, Power BI, Tableau | **Languages:** Python, SQL, R

Certifications

AWS Certified Cloud Practitioner — *Issued May 2025*

Education

Boston University — Boston, MA

B.S. in Data Science | Dean's List: 2022 - 2025 | DS Undergraduate Tutor (Mar 2023 - Dec 2024) — *09/2021 – 01/2025*