Rohan Anand

anandro@bu.edu | linkedin.com/in/rohan-h-anand | (603) 233-4670

Experience

Data Engineer — Mar 2025 – Present

Dataeconomy — *Charlotte, NC*

- Developed a recommendation system for a data marketplace using Alternating Least Squares with Implicit Feedback to predict user preference for datasets
- Increased precision for top-k items by 50% compared to popularity baseline through hyperparameter tuning using Optuna
- Operationalized model by containerizing it with Docker and deploying it on AWS ECS
- Created an endpoint using FastAPI to serve predictions to users
- Orchestrated a CI/CD pipeline using Jenkins to automate the model retraining, validation, and deployment process
- Implemented data quality profiling and validation service using Snowpark in a Snowflake ETL pipeline, allowing calculation of row and column level metrics, ensuring data integrity
- Added functionality to segregate data into valid/invalid sets and automatically generate detailed data quality reports
- Incorporated feature to load Snowflake-managed Iceberg tables from S3 and apply SCD types 0-4 for historical data management
- Rewrote internal data validation package using Great Expectations in a PySpark ETL pipeline, adding new quality checks and categorization of constraints, reducing runtime by 70% on large datasets
- Engineered a low-latency pipeline using Snowpipe to ingest real-time data from Kafka into Snowflake
- Architected a FastAPI application for a no-code Airflow orchestration service with connections to third-party services such as Talend, Informatica, Confluence, Teams, etc., using GraphQL for front-end ingestion and its REST API for backend execution

AI Engineer — Sep 2024 – Jan 2025

BU Spark! — *Boston, MA*

- Developed a RAG-LLM using OpenAI's API for the National Lawyers Guild to identify discrepancies in punitive outcomes for police officers by severity, complete with reasoning based on allegation and description of events, reducing investigation time by 80%
- Extracted officer information from historical spreadsheets from 2011 - 2024, generated embeddings and inserted into vector database using ChromaDB
- Utilized advanced techniques such as prompt engineering, chain-of-thought reasoning, feature augmentation with past offenses in LangChain to increase explainability and accuracy by 40%
- Visualized differences between AI-recommend and actual outcomes, focusing on variations across officer ranks and time spent in department
- Implemented full-stack functionality using FastAPI and JavaScript to serve LLM outputs, complete with CSV upload for future records
- Selected as top project and presented to client and data science faculty

Data Services Intern — May 2024 – Aug 2024

Axis Technology, LLC — *Boston, MA*

- Developed a Python script to populate CRM database with a variety of realistic fake personal information for a ML model that detects private information, ensuring robust training data, increasing validation accuracy by 60%
- Implemented flexible data generation features to scale number of training samples from 1k to 100k+, while maintaining similar runtime
- Created a script to label to parse JSON of database schema and label column information type to augment features for model
- Formulated test code to validate OpenSearch produced relevant results in a program that identified similar tables in a database

Data Analyst Intern — Jun 2022 – Aug 2022

AS Insurance Agency — *Manchester, NH*

- Concatenated personal information from 1,000+ customers and corresponding insurance statements from 1,500+ declaration pages using Pandas
- Constructed SQL queries on Snowflake to identify key customer segments and developed 10-15 Tableau dashboards to target customers for renewal
- Streamlined renewal process, leading to 95% retention in clients

Projects

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

- Developed and maintained a database of 2.7M+ Boston 311 service requests over 12 years, automating daily updates via API integration, to analyze community service equity
- Created interactive graphs using ipywidgets to identify trends in request volume, submission sources, and resolution times across neighborhoods
- Constructed a PowerBI map visualization encoded with social vulnerability index data to highlight leading requests by geographic area
- Produced and presented a PowerBI report analyzing disparities in request types and resolution times across income levels, presented findings to client and data science faculty

Technical Skills

Languages: Python, SQL, R | **Cloud/Platforms:** Snowflake, AWS, Azure, Docker, Kafka, Airflow, Pinecone, PostgreSQL, Jenkins, Talend, Informatica, Confluence, Jira, Postman | **Libraries/APIs:** Pandas/NumPy, Scikit-learn, FastAPI, Optuna, OpenAI, Snowpark, Pytorch, PySpark, LangChain, ChromaDB, Great Expectations, Opensearch | **Tools:** Git, Tableau, PowerBI, dbt

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*