Rohan Anand

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

Experience

Data Engineer March 2025 – Present

Dataeconomy *Charlotte, NC*

- Developed a personalized recommendation system for a data marketplace by adapting Alternating Least Squares to process implicit feedback from time-series user interaction data, training on historical usage to predict dataset preferences
- Optimized model performance by conducting hyperparameter tuning with cross-validation using Optuna
- Built a data quality profiling and validation service for Snowflake datasets, leveraging Snowpark to implement comprehensive row and column-level metrics, offering functionality similar to PyDeequ
- Engineered data quality service to segregate data by validation status into passed/failed datasets and generate reports detailing outcomes and record counts
- Integrated metrics into the existing application architecture using FastAPI endpoints
- Implemented functionality using Snowpark to generate Snowflake-managed Iceberg tables from S3 data and metadata, enabling application of Slowly Changing Dimension (SCD) Types 0-4 for robust historical data tracking and enhanced data management
- Engineering a low-latency data pipeline with Snowpipe Streaming's high-performance architecture to ingest real-time event data from a Kafka cluster into Snowflake

Data Services Intern May 2024 – August 2024

Axis Technology, LLC *Boston, MA*

- Developed a Python program that generates realistic personal information, enabling the team to efficiently train an ML model to detect and obscure sensitive data
- Implemented flexible data generation features to control test data volume, helping to identify gaps in model performance
- Created a program that processes database structures formatted in JSON to automatically label columns, improving model accuracy
- Formulated test code to validate OpenSearch functionality in a ML model comparing metadata between tables

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
- Provided quotes for 5-10 customers per week with expiring insurances, retaining over 95% of clients

Projects

Police Misconduct Data Analysis for NLG Oct. 2024 – December 2024

- Analyzed internal investigation findings (2011-2020) for Massachusetts National Lawyers Guild using statistical methods and machine learning techniques
- Developed a RAG LLM system using OpenAI's API to recommend objective punitive outcomes based on allegations and descriptions, trained on Boston Police Department rules
- Visualized differences between AI-recommended disciplinary actions and actual outcomes, focusing on variations across officer ranks and past disciplinary records
- Implemented advanced sentiment analysis to classify misconduct severity, and created interactive maps depicting relationships between officer allegations and median household income
- Selected as top project to present findings to data science faculty and client

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, present findings to data science faculty and client

Modern Data Pipeline for Analyzing Covid-19 Jan. 2023 – May 2023

- Analyzed Covid-19 spread and policy effectiveness across 10 countries using Azure Data Factory for ETL and PowerBI for visualization
- Integrated diverse data sources (JSON, Parquet) from Azure Data Lake Storage into a unified fact table within a star schema design using SQL transformations
- Created interactive visualizations in PowerBI to identify trends and compare the impact of different policy implementations

Technical Skills

Languages: Python, SQL, R | **Cloud/Platforms:** Snowflake, AWS, Azure, Docker, Kafka, Airflow, Pinecone | **Libraries/APIs:** Pandas/NumPy, Scikit-learn, FastAPI, Optuna, OpenAI, Snowpark, Pytorch, Pyspark (Apache Spark), LangChain, ChromaDB, | **Tools:** Git, Tableau, PowerBI

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*