Rushil Jagat Sheth

sheth.rushil@gmail.com | rushilsheth.com | (314) - 766 – 3990 | San Francisco

Education

University of San Francisco San Francisco, CA **June 2020**
Master of Science in Data Science

University of California – Berkeley Berkeley, CA **May 2017**
Bachelor of Arts in Applied Mathematics
Bachelor of Arts in Statistics

Personal Projects

Knowledge Base Chatbot - allow users to ask any question related to internal documentation

- Optimized the scraping, reading, embedding and storage of multimodal documents via parquet files, Ray, and distributed computing in conjunction with Pinecone. Deployed and hosted on Anyscale via Ray and FastAPI
- Evaluated retrieval and quality of the RAG pipeline with empirical measures of quality, cost, and latency
- Flexible architecture to allow the agent to utilize various search tools: elasticsearch, vector store, Neo4j graph db, or the internet

Work Experience

Machine Learning Engineer Lead **February 2022 – Present**
EchoAI San Francisco, CA

- Developed and implemented generative algorithms via LLMs from 0→1, ensuring autonomous, precise, streamlined insights generation through cutting edge research, RLHF, LangChain, and vector databases
- Directed technical projects autonomously, meticulously scoping requirements and sequencing work for the engineering team, and serving as the primary LLM and ML expert advisor to the CTO and team
- Contributed to the creation of an external service using AWS Lambda for large-scale machine learning models. This effort focused on offloading computational tasks to enhance web app performance, and involved the integration of pgvector as the embedding database for optimal LLM accuracy and efficiency
- Hosted and fine-tuned open source LLMs via LoRA and DSPy for cost reduction and flexibility on task routing. Utilized a dataset, both human and synthetic, that was aligned with customer intent
- Collaborated with teams across departments leveraging dbt and Hex for advanced analytics and automated the processing of customer data requests, boosting satisfaction as the sole Data personnel at the company

Data Scientist **July 2020 – February 2022**
eHealth San Francisco, CA

- Identified customers likely to churn through a Random Forest classifier and MLflow for model selection, which resulted in a high AUC and recall. Results were made accessible via an ETL pipeline to Snowflake
- Determined lifetime value (LTV) for customers using a Cox Proportional Hazard model. Delivered results via AWS SageMaker and Lambda to create a REST API
- Created an end-to-end model, pipeline, and monitoring to predict the LTV associated with sending a mailer to an individual via Spark, XGBoost and H2O which led to over 80% LTV with a 50% reduction in cost

Data Science Intern **October 2019 – June 2020**
New York Mets Remote

Business Analyst **September 2018 – July 2019**
Pinterest San Francisco, CA

Proficient Technologies: Python, LLMs, Ray, PyTorch, Distributed Computing, Computer Vision, LLMOps, SQL, Next.js, MLOps, Go, Hex, dbt, AWS, FastAPI