Sulman A. Khan

1047 East 14th St., Apt #2, Brooklyn, NY 11230
(718) 755-6739 • sulman@vt.edu • www.linkedin.com/in/sulman-khan • https://sulmank.github.io

Skills

Programming languages: Python, SQL (PostgreSQL, MySQL), JavaScript, HTML/CSS
Cloud Platforms: AWS, GCP
Tools & Frameworks: Git, Docker, Terraform, Kubernetes, Apache Airflow, PyTorch, Flask, LangChain
Data Tools: BigQuery, Looker, Apache Spark, dbt
Domains: A/B Testing, Machine Learning, Deep Learning, Natural Language Processing, Recommender Systems, Generative AI, Prompt Engineering, Retrieval-Augmented Generation

Work History

Nuage Software Corporation, NY **June 2024 - Present**
AI/DS Consultant

- Engineered a scalable ETL pipeline on GCP, orchestrating workflows with Apache Airflow, transforming data with dbt and Apache Spark, and optimizing storage in BigQuery to support customer churn prediction and retention strategies.
- Implemented customer churn prediction models, applying advanced feature engineering to drive actionable insights and reduce churn.
- Consulted on generative AI deployment, leveraging retrieval-augmented generation and prompt engineering to develop tailored solutions for complex business challenges.

Fingercramp, NY **May 2018 - January 2024**
Data Scientist

- Applied decision-based heuristics on player match statistics to develop a character balancing model, effectively doubling the number of characters utilized from 16 to 32 and improving overall game balance.
- Established and maintained a PostgreSQL database for match statistics, enabling complex querying across multiple tables and schemas for in-depth data analysis and reporting.
- Designed data visualization dashboards for a streaming platform, driving a 48% increase in peak viewership by improving user engagement.

Personal Projects

Reddit AI Pulse: Building an AI-Powered Data Pipeline **Winter 2024**
https://github.com/SulmanK/reddit_ai_pulse_cloud_public

- Designed and implemented a scalable ETL pipeline on GCP, using Apache Airflow for orchestration, GCS as a data lake, dbt and Apache Spark for data transformations, and BigQuery for warehousing. Automated daily processing of AI-focused subreddit data, ensuring reliability with Grafana, Prometheus, MLflow, and Cloud Monitoring.
- Deployed NLP models (RoBERTa for sentiment analysis, BART for text summarization, Gemini AI for content analysis) in a React-based app with dynamic visualizations, delivering categorized insights and actionable summaries of Reddit discussions.

RecSys Challenge 2024 **Fall 2024**
https://github.com/SulmanK/2024-Recsys-Challenge

- Engineered and preprocessed the large-scale EB-NeRD dataset for the Ekstra Bladet RecSys Challenge 2024, integrating user interaction logs, session metadata, and enriched article features to prepare high-quality data for machine learning models.
- Applied and optimized personalized news recommendation models for click-through rate prediction, leveraging feature selection, recommendation techniques, and ROC-AUC evaluation. Achieved a top 100 placement in the Ekstra Bladet RecSys Challenge 2024.

eBay: Phone Auction Aide **Fall 2020**
https://github.com/SulmanK/eBay-web-crawler-phone-auctions

- Created a Python-based web scraper to gather phone auction data from eBay, deploying workflow services on AWS for automation and maintenance of a PostgreSQL database, which resulted in a 25% improvement in efficiency and throughput.
- Launched a user-friendly application for real-time monitoring of phone auctions, providing valuable metrics to assist in auction selection, and improving the user experience.

Education

Stony Brook University, Stony Brook, NY **May 2018**
Masters of Science, **Electrical Engineering (Concentration in Machine Learning Systems)**

Virginia Polytechnic Institute and State University, Blacksburg, VA **May 2016**
Bachelors of Science, **Materials Science and Engineering**