

Keyush Shah

+1 (267) 278-1269 | keyush.shah12@gmail.com | [in](#) [LinkedIn](#) | [Website](#) | [Blog](#)

EDUCATION

University of Pennsylvania (School of Engineering and Applied Science)

Philadelphia, PA

Master of Science in Engineering (MSE) in Data Science; GPA 3.9/4.0

Aug 2023 – May 2025

Coursework: Database & Information Systems, Adv. Deep Learning (Diffusion models, LLMs, Multimodal architectures), Generative Modeling, Adv. Machine Perception, Machine Learning, Statistical/Probability models in Marketing

Kirori Mal College (KMC – University of Delhi)

Delhi, India

Bachelor of Science (BS) in Statistics; GPA: 8.11/10

July 2017 – Aug 2020

EXPERIENCE

Research Intern, [Computational Social Listening Lab \(UPenn\)](#)

May 2024 – Present

Project 1: (Misinformation)

- Worked on identifying misinformation on social media and whether it relates to health outcomes segmented by race.
- Extracted linguistic features from posts using [DLATK](#) and applied LDA for topic modeling and performed correlation analysis.
- Built NLP pipelines to detect health-related misinformation from ~20K social media posts using RoBERTa and LDA topic modeling; improved **classification precision by 20%** through alignment-based entailment.
- Unified data from various online survey platforms (**10K+ responses**) in a secured server via MySQL and performed feature engineering in Pandas to prepare data for further downstream tasks.

Project 2: (IH Risk Model)

- Developed a pipeline to predict Incisional Hernia (IH) risk from unstructured operative notes and structured EHR data in a surgical cohort of 10k+ patients.
- Engineered a few-shot GPT-based extraction pipeline to label operative features (e.g., incision, ostomy) from redacted notes **and reduced noisy extractions by 30% vs. BERT embeddings**.
- Scaled the pipeline for 10k+ notes using **parallelized batch inference** and memory-aware data chunking.

Data Science Intern, [Universal Media \(PA, USA\)](#)

May 2024 – Aug 2024

- Led the development of **3+ data pipelines** using Azure Data Factory (ADF), facilitating the seamless ingestion and transformation of diverse data sources into the Azure environment.
- Developed python scripts for data transformation, stored them in Blob storage and executed them via batch activity in ADF.
- Drove product marketing insights by building **Mixed Media models (MMM) in Azure Synapse**, analyzing marketing channel impacts on media diversity metrics. Built **Power BI** dashboards to deliver actionable insights for optimizing client strategies.
- Authored **5 stored procedures** in SQL, automating repetitive tasks and improving query performance by over 30%.

Assistant Manager, [IIFL Finance Ltd](#)

Apr 2022 – July 2023

- Analyzed ETL process failures and created **10+** paginated reports in SSIS to help the management track 1000+ branches.
- Optimized & migrated complex SQL queries from an obsolete database server that improved the **reporting services by ~40%**.
- Digital Adoption:** Led a product-focused initiative to identify digitally savvy customers by engineering features and building ADF pipelines to track campaign behavior. Trained a Random Forest model (**with a 90% accuracy**) in Azure ML Studio to score customer readiness, enabling personalized outreach and boosting digital **disbursal adoption by 50%**.

SELECTED PROJECTS

- Diffusion Transformer (2024):** Implemented PatchVAE with convolutional encoders and patch-based decoding for fine-grained feature extraction. Trained a Diffusion Transformer to sample from the latent space of PatchVAE, achieving a **30% reduction** in FID score and **2x greater feature diversity** compared to VAE-generated samples. [\[Github\]](#)
- Statistical Segmentation (2024):** Modeled customer purchase behavior using Pareto II and Weibull distributions; optimized parameters using Excel Solver and back tested on historical unit sales, improving segment alignment with **actual revenue by 21%** and informing future campaign targeting. (Solver, Marketing). [\[Link\]](#)
- FitBit (2024):** Engineered a Django health chatbot leveraging PostgreSQL for robust patient data management, featuring an LLM-agnostic architecture with seamless model switching via **Langchain that reduced overhead by 40%**. Optimized memory usage for long conversations through entity extraction, and **selectively triggered LLM responses** to reduce latency. [\[Github\]](#)
- Real Estate Prediction (2023):** Built an end-to-end ML pipeline using XGBoost, achieving 96% accuracy in house price prediction. Developed a Flask API to serve predictions and integrated it with a front-end (HTML, CSS, JavaScript). [\[Github\]](#)

TECHNICAL SKILLS

Programming Languages: Python, C/C++, SQL, R programming, JavaScript

Frameworks: PyTorch, React, NodeJS, MongoDB, HTML/CSS, Neo4J, OpenCV, Apache Spark, MLOps, PySpark, Django, Flask

Platforms & Tools: AWS, SSMS, Docker, Airflow, Azure Data Factory, Azure DevOps, A/B testing, Power BI, SPSS