Saransh Surana

+1 (631) 816-8344 | ssurana818@gmail.com | linkedin.com/in/saransh-surana | saransh-surana

EDUCATION

SUNY - Stony Brook University **Aug 2023 - May 2025**
Master of Science in Data Science *Stony Brook, NY, USA*

- **Research Assistant (Jan 2025 – May 2025)** - Built Python pipelines for unstructured PDFs (OCR + cleaning).

SKILLS

- **Languages:** Python (primary), SQL, TypeScript, Java, C/C++; Linux/Bash; data structures & algorithms
- **ML & DL:** PyTorch, TensorFlow/Keras, scikit-learn, Numpy, pandas; XGBoost, LightGBM; NLP(SpaCy), OpenCV, spaCy
- **LLM / GenAI:** prompt engineering; Gemini Pro + Transformers; agents/RAG workflows; LLM architecture + operations; tool/function calling; guardrails + structured outputs (JSON Schema/Pydantic)
- **Retrieval/RAG:** Sentence-Transformers, FAISS (VectorDB); chunking, hybrid search, reranking; LangChain/LlamaIndex
- **Data & MLOps:** Spark, Kafka, BigQuery; FastAPI; Docker, Kubernetes, CUDA; CI/CD; Agile; Airflow/MLflow; monitoring (Prometheus/Grafana); AWS/GCP/Azure

WORK EXPERIENCE

Schizophrenia & Psychosis Action Alliance **Jul 2025 – Present**
AI Software Engineer Remote, USA

- Developed a distributed **LLM extraction pipeline** using **FastAPI**, **Redis**, and **asyncio** to transform 3,200 county-level JSON files into structured housing datasets.
- Implemented **parallel inference, caching, and retry logic** with Redis queues to sustain multi-county **Gemini Flash** processing and reduce redundant API calls.
- Built **prompt-templated modules** for relevance filtering and **schema-aligned JSON** extraction.

Ford Motor Company **May 2024 – Aug 2024**
Data Science Intern Dearborn, MI, USA

- Developed + evaluated **unsupervised models (Isolation Forest + PCA, One-Class SVM)** with metric-driven iteration; **78% recall, 73% precision**.
- Partnered with engineering and business stakeholders to translate model results into actionable decisions.

Napuor **Aug 2022 – Jan 2023**
Data Science Intern Bengaluru, KA, India

- Trained and deployed **time-series demand-forecasting** and inventory-optimization ML models using **XGBoost** on GCP, reducing forecast error by 18%.
- Designed ETL pipelines (Spark + SQL): Hive + Kafka; unified millions events/month; OLAP + sub-sec inference.
- Developed, built, and maintained scalable ML pipelines for training, evaluation, experiment tracking, and versioned artifacts using **MLflow** and **Prometheus/Grafana**.

PROJECTS

Multimodal Meeting-to-Storyboard System | *React.js, FastAPI, Python, Whisper.cpp, Stable Diffusion XL, PyTorch, Vite*

- Developed an agent-based multimodal pipeline converting meeting audio → text → storyboard images using Whisper.cpp, Transformers, and SDXL.
- Optimized CUDA-based TensorRT inference to process 20-minute audio in <2 minutes with batched GPU execution.
- Containerized modular React + FastAPI services using Docker for collaborative, low-latency generation workflows.

Recommendation System (RL) — Book Recommender | *Python, PyTorch, Stable-Baselines3, Gymnasium, Tensorflow*

- Trained a PPO agent on 10K+ interactions, achieving 19% higher reward than BiLSTM and logistic baselines.
- Iteratively tuned models in ambiguous problem settings, balancing performance trade-offs and stakeholder needs.

RAG-based Financial Document Assistant | *Python, LangChain, FAISS, OpenAI API*

- Built a containerized LangChain + FAISS RAG service for querying financial documents with structured outputs.
- Reduced hallucinations by 25% via semantic chunking and FAISS-based re-ranking.

OPEN SOURCE CONTRIBUTIONS

- **Haystack Core Integrations** (deepset-ai): Two merged PRs - fixed llama-stack integration and CI workflow; streamlined CI for archived Google integrations.
- Contributed doc fixes and features to **Skrub** and **Statsmodels** (merged into main).
- Merged parametrized Transformers smoke test in **Outlines** (dottxt-ai) for tokenizer robustness.

TECHNICAL APPENDIX

OTHER PROFESSIONAL EXPERIENCE

Women In Science and Engineering (Stony Brook University) **Mar 2024 – Apr 2024**
Machine Learning Instructor *Stony Brook, NY, USA*

- Led hands-on sessions explaining end-to-end machine learning workflows, including algorithm selection, feature engineering, and evaluation using Python-based implementations.

The Sparks Foundation **Apr 2021**
IOT and Computer Vision Engineer Intern *Remote, India*

- Implemented an object-detection pipeline using OpenCV DNN with SSD MobileNet (COCO), performing real-time image and video inference with confidence thresholding, bounding-box rendering, and class-label mapping.

Geeks for Geeks **Nov 2020 – Mar 2021**
Technical Writer (Java) Intern *Remote, India*

- Authored Java-focused data structures and algorithms articles, emphasizing optimized solutions, edge cases, and interview-oriented explanations for a large learner audience.

Suven Consultants & Technology Pvt. Ltd. **Oct 2020**
Java Coding Intern *Remote, India*

- Built Java console applications including a Consumer Loan Assistant and Home Inventory Manager, applying core OOP principles, modular design, and problem-driven development.

ADDITIONAL TECHNICAL PROJECTS

Employee Attrition Prediction | *Python, Logistic Regression, SHAP*

- Modeled attrition risk on the IBM HR Analytics dataset using logistic regression with regularization and class balancing, applying statistical modeling principles.
- Achieved 0.86 AUC and documented rationale for model choice using SHAP-based feature importance.

E-Commerce Platform | *React.js, Node.js, Express.js, MongoDB, REST APIs, GCP*

- Engineered a full-stack MERN app with product search, filters, JWT authentication, and secure Razorpay checkout.
- Deployed microservices on Google Cloud (GKE, Cloud Build, Cloud Storage) with CI/CD, using AI tools with MCP style data fetching (logs, metrics, API responses) to speed debugging and development.

Backend Development for Ad Reporting System | *Python, PostgreSQL, OpenSearch, Kafka*

- Developed backend services using relational (PostgreSQL) and NoSQL (OpenSearch) stores for fast data retrieval.
- Implemented reporting workflows on PostgreSQL and integrated OpenSearch for low-latency log search and filtering.

ACHIEVEMENTS

- Placed **Top 14% (Rank 376/2722)** in a Kaggle machine learning competition, demonstrating competitive model development, feature engineering, and evaluation.
- Achieved **98.21 percentile** in the **Common Admission Test (CAT)**, a highly competitive national exam in India emphasizing quantitative aptitude, data interpretation, and logical reasoning.

PUBLIC PROFILES

- **Kaggle:** kaggle.com/saranshsurana07
- **Medium:** medium.com/@saranshsurana