SWATI

646-729-6173 swati.swati@stonybrook.edu LinkedIn GitHub

PROFESSIONAL SUMMARY

- Results-driven **Data Scientist** and **Software Engineer** with 5+ years of enterprise experience building scalable ML systems and cloud infrastructure. Proven expertise in **machine learning**, **deep learning**, and **DevOps** practices, with a track record of deploying production-grade solutions that drive significant business impact. Specialized in **time-series forecasting**, **natural language processing**, and **distributed systems** using **Python**, **PySpark**, **AWS**, and **Kubernetes**. Currently pursuing Master's in Data Science with focus on **generative AI** and **multilingual semantic search** systems.

EXPERIENCE

AI Software Engineer Intern — May 2025 – August 2025

Voltihost LLC — Florida, United States

- Built **XGBoost** regression models for demand forecasting achieving **85% R^2 accuracy** and **7.2 MAE** while handling **121-day** dataset with **66.9%** missing data through advanced interpolation strategies
- Engineered comprehensive feature pipeline creating **15+ features** including time-series lags (**3, 5, 7 days**), rolling statistics, weather correlations, and temporal indicators for **ML model training**
- Implemented advanced data preprocessing handling missing values, outlier detection using **IQR methodology**, and automated **feature engineering** pipeline for real-time inference
- Constructed production **ML infrastructure** exporting models to **ONNX format**, implementing model serialization with metadata management, and deploying **live forecasting API** with configurable horizons (**1-30 days**)
- Maximized model performance through **hyperparameter tuning**, **cross-validation** strategies, and **feature selection** achieving improved accuracy over baseline models (**Prophet** R^2 = -229 vs **XGBoost** R^2 = 0.85)

Senior Software Engineer — August 2019 – July 2024

Infosys Ltd. | Awards: Insta Award, Constant Contributor Badge — Pune, India

- Installed Finacle Core Banking on **OpenShift** and **AWS** clusters, establishing and managing environments for **20+** applications, with **5** installations annually to address new functional business requirements and updates
- Built and deployed **JFrog** and **ECR images** in **Kubernetes** namespaces and **AWS**, automating patch deployments and managing deliverables, resulting in a **50%** reduction in deployment time and a **30%** increase in deployment accuracy
- Deployed and managed Dockerized applications on **AWS EKS**, using **Kubernetes**, **Docker**, **Ansible**, and **GitLab** to create **15+ CI/CD** pipelines, achieving a **40%** reduction in infrastructure costs through optimized resource allocation and scalability
- Led a DevOps team of **3** in managing end-to-end development and operations for projects in **SEPA** and **APAC** region, overseeing CIT, E2E, and production environments for **50+** integrations, driving **$3M** in annual revenue for client (BOFA)
- Hosted and managed **Escrow** deposit deployments for Finacle applications across **3+** versions with NCC Group in the UK region for **Bank of America** as a third-party client
- Maintained **OLAP data** and production environment reliability for Finacle **e-banking** by replicating and creating **50+** databases across multiple accounts to support application structure changes in the APAC region
- Governed and enhanced **10+** cloud servers by configuring **crontab jobs**, archiving logs, ensuring efficient disk management, and using **Jenkins CI/CD** pipelines for source code deployment, backups, and memory alerts
- Generated **20+** detailed reports with iReport, optimizing backend processes through complex **SQL** and **PL/SQL** queries
- Enhanced Finacle's functionality for an investment banking firm by customizing **100+ web pages** and menus using **JavaScript** and **JSPs**

SKILLS

- **Programming Languages**: **Python**, **R**, **C++**, **Java**, **Shell Scripting**, **Bash**, **SQL/PL-SQL**, **Linux**, **UNIX**
- **Data Science & Machine Learning**: **Pandas**, **NumPy**, **Exploratory Data Analysis (EDA)**, **Statistical Learning**, **Statistical Computing**, **Deep Learning**, **Probability**, **Scikit-Learn**, **PyTorch**, **Tableau**, **Matplotlib**, **Seaborn**, **XGBoost**, **Random Forest**, **SVM**, **Decision Trees**
- **Cloud & DevOps**: **Vercel**, **Kubernetes**, **Docker** (containerization), **GitHub Actions**, **GitLab**, **Git** (version control), **Ansible**, **AWS** (**EKS**, **EC2**, **S3**, **AWS SageMaker**), **Airflow**, **Hadoop**, **HDFS**, **OpenShift**, **Jenkins**, **Tortoise SVN**, **JIRA**, **WinSCP**, **RabbitMQ**
- **Databases & Data Warehousing**: **Relational DBMS**, **Excel**, **Oracle** 11g/12c/19c, **MySQL**, **Supabase**, **Data Warehousing**, **Weaviate** (Vector DB)
- **AI/ML Frameworks & Tools**: **FastAPI**, **LangChain**, **PySpark**, **Apache Kafka**, **ONNX**, **EmbeddingGemma**, **OpenWebUI**, **Salesforce BLIP**, **Qwen 2.5 VLM**, **Streamlit**
- **AI Development Tools**: **Cursor**, **Claude**, **Lovable AI**, **vo by Vercel**

EDUCATION

Master of Science in Data Science — **August 2024 – May 2026**
Stony Brook University (SUNY), New York — GPA: 3.40/4.00

Bachelor of Technology in Computer Science — **July 2015 – May 2019**
Chitkara Institute of Science and Technology, India — GPA: 8.97/10.00

PROJECTS

Cross-Lingual Document Retrieval System with Multimodal AI

(Python, Docker, FastAPI, LangChain, EmbeddingGemma, Weaviate, Open WebUI, Salesforce BLIP, Qwen 2.5 VLM)

- Architected production-grade multilingual semantic search engine supporting **English**, **Hindi**, and **Spanish** queries, achieving **92%** retrieval accuracy using **EmbeddingGemma** models and **Weaviate** vector database
- Built end-to-end **MLOps pipeline** with **FastAPI** and **Docker**, deploying scalable **RESTful API** for embedding generation, vector indexing, and evaluation metrics
- Developed interactive **chat interface** using **OpenWebUI** with **multimodal capabilities** via **Salesforce BLIP** and **Qwen 2.5 VLM** for image-text understanding

Netflix Recommender System at Scale

(Python, PySpark, Spark MLlib ALS, MPI, Slurm, SeaWulf HPC, NeuMF, Parquet, Numba, PyTorch, AWS/GPFS)

- Engineered end-to-end **Netflix Prize** recommender on **100M+ ratings**, converting raw data to columnar **Parquet** and orchestrating distributed preprocessing with **PySpark** across **SeaWulf HPC** nodes
- Built scalable collaborative filtering with **Spark MLlib ALS** and optimized matrix factorization (**FunkSVD**) using **MPI** and **Numba**-accelerated SGD; managed long-running jobs via **Slurm**
- Implemented deep learning **NeuMF** (GMF + MLP) in **PyTorch** with mixed precision and checkpointing; tuned on multi-GPU HPC, tracked metrics/plots in project repo NetflixRecommenderSystem

Real-Time Demand Forecasting with Advanced ML Pipeline

(Python, XGBoost, Scikit-learn, Pandas, NumPy, ONNX, FastAPI)

- Engineered demand forecasting system achieving **85% R^2 accuracy** and **7.2 MAE** on 121-day dataset with **66.9% missing data** through advanced interpolation and **15+ engineered features**
- Implemented production-ready **ML infrastructure** with **ONNX format** serialization and deployed **live forecasting API** supporting configurable horizons (1-30 days) with sub-second latency
- Optimized model through **hyperparameter tuning** and **cross-validation**, outperforming baseline (**Prophet** R^2 = -229 vs **XGBoost** R^2 = 0.85)

NYC Crime Pattern Analysis & Predictive Modeling

(R, Statistics, Machine Learning, EDA, Data Visualization)

- Conducted borough-level crime pattern analysis using **Random Forest**, **SVM**, and **Decision Trees**, achieving **97.24% test accuracy** with extensive **EDA** and feature engineering
- Developed predictive models identifying high-risk areas and temporal patterns with actionable insights through advanced **data visualization**

Enterprise Fitness Center Database Management System

(Python, MySQL, Streamlit, SQL/PL-SQL)

- Designed comprehensive **relational database system** with **10+ entities** and **45+ tuples** using **EER diagrams** and normalization principles
- Developed interactive **Streamlit** web application with dynamic query interface for real-time data insights and business intelligence dashboards