Samhitha Nuka

Data Scientist & Machine Learning Engineer

samhithanuka51@gmail.com | +1 812-361-5059 | LinkedIn | Github | Portfolio

PROFESSIONAL SUMMARY

Data Scientist & ML Engineer with over 2 years of experience in developing scalable ETL pipelines, deploying Generative AI applications, and enabling real-time analytics on AWS, Azure, and Databricks. Hands-on experience managing datasets with over 7 million records and achieving low-latency responses under 200ms with LLMs. Focused on integrating machine learning models into production, streamlining workflows, and delivering actionable insights through dynamic dashboards and automation.

EDUCATION

MS in Data Science, *Indiana University Bloomington* 08/2023 – 05/2025 | Bloomington, United States

SKILLS

Machine Learning: LSTM, CNN, Transformers, Random Forest, SVM, K-Means, PCA, SMOTE, NLP, Model Evaluation

Generative AI & LLMs: Gemini, Gemma, Groq, ChatGPT, LLaMA, LangChain, FAISS, RAG, Prompt Engineering, Agentic AI

Data Engineering: ETL/ELT (Databricks, Spark, Informatica), Kafka, Hadoop, Airflow, Snowflake, Redshift, BigQuery, MongoDB, Cassandra

Visualization: Power BI, Tableau, Looker Studio, Excel (Pivot Tables, Power Query, VLOOKUP), Matplotlib, Seaborn, SAS

Programming: Python, SQL, R, Java, Bash, JavaScript, PHP, HTML, CSS

Cloud & Tools: AWS (Lambda, S3, SageMaker), GCP (BigQuery, Cloud Storage), Azure (Databricks, DevOps), Docker, Git, Postman

WORK EXPERIENCE

Indiana University Bloomington, *Graduate Research Assistant* 01/2025 – present | Bloomington, United States

- Constructed a Python pipeline to process and label over **10,000 time-series events**, preparing data for classification based on temporal sequence patterns.
- Extracted and transformed **15+ statistical features** such as inter-event intervals, rolling variance, and delta rates, and trained Random Forest, SVM, and Logistic Regression models, achieving 92% accuracy with **5-fold cross-validation** and **GridSearchCV.**
- Modeled temporal trends using linear regression, yielding an **average R^2 of 0.87** and identifying timing anomalies through residual analysis.
- Created 10+ visualizations with Matplotlib and Seaborn to explain model performance, feature contributions, and temporal behavior to non-technical stakeholders.

Cognizant Technology Solutions, *Junior Machine Learning Engineer* 08/2021 – 07/2023 | Chennai, India

- Catalyzed the integration of OpenText ECM with AWS SageMaker, creating **42+ RESTful APIs** to automate data extraction and analysis, establishing a reusable integration standard.
- Enhanced legacy systems to support ML workflows, resulting in a **30% improvement in processing efficiency** and a **40% boost in search index speed**.
- Orchestrated **serverless inference endpoints** using API Gateway and Lambda, sustaining less than **100ms latency**, **99.9% uptime**, and handling **500+ concurrent requests daily** for reliable ML pipeline execution.
- Implemented real-time monitoring with Amazon CloudWatch and automated deployment with AWS CodePipeline, accelerating model release cycles and ensuring **production-grade MLOps**.

ACADEMIC PROJECTS

Intelligent Portfolio Rebalancing Using Reinforcement Learning, *Proximal Policy Optimization (PPO) | Agent | Agentic AI*

- Designed a custom OpenAI Gym environment and trained a **PPO agent** on historical S&P 500 data to enable dynamic asset rebalancing.
- Achieved **~18% higher Sharpe ratio** vs. SPY benchmark with daily reallocation strategy.

Blog Generation using Amazon Bedrock and LLaMA, *Amazon Bedrock | LLaMA 3 70B Instruct | AWS Lambda| API Gateway | Amazon S3*

- Deployed a serverless blog generator with Lambda, API Gateway, and Bedrock (LLaMA 3 70B), achieving less than **200ms latency**.
- Leveraged **Postman and S3** to enable prompt-response traceability and scalable **LLM deployment**.

Scalable Data Pipeline Development for Real-Time Traffic Analytics, *Azure Databricks | SparkSQL | CI/CD | ETL*

- Engineered a real-time ETL pipeline handling **100,000+ traffic records/day**, using Spark SQL and Medallion Architecture on Azure.
- Designed and deployed 3 Power BI dashboards that improved decision-making speed by **30%**.
- Automated deployments via Azure DevOps, reducing manual deployment time by **80%**.

End-to-End AI Application Suite for Multimodal Interaction, *Streamlit | Gemini Pro/Flash | FAISS | Groq | LLM*

- Created 4 AI-powered Streamlit apps for Q&A, PDF/image analysis, and SQL generation using Gemini Pro/Flash and SQLite.
- Integrated Groq and Gemma for **30–50ms** responses and **FAISS** for sub-second search across **30+ documents.**

Streaming Cybersecurity Analytics using Kafka, Spark, and S3, *Cybersecurity analytics | Containerization | Real-time data ingestion*

- Architected a real-time data pipeline with **Kafka**, **Spark**, and **Docker**, streaming **10K+ logs/min** to **AWS S3** for scalable storage and analysis.
- Enabled automated intrusion detection by integrating a **Random Forest model** in Spark, achieving **~99% accuracy** on the UNSW-NB15 dataset.

Credit Card Fraud Detection with PCA, SMOTE, and Ensemble Learning, *PCA | SMOTE | XGBoost | Ensemble Learning*

- Built fraud detection pipeline achieving **95.6% F1-score**, using PCA and SMOTE on imbalanced datasets.
- Compared Random Forest, XGBoost, and Logistic Regression, deploying a stacked model that Increased **recall by 12%**.

CERTIFICATIONS

- AWS Machine Learning Engineer – Associate
- Machine Learning offered by Stanford University
- Databricks Data Engineer Associate
- Rest API (Intermediate) assessment in HackerRank
- Google Data Analytics Professional Certificate
- OpenText Content Server Developer v20.4
- IBM Python for Data Science and AI