Samhitha Nuka

Data Scientist & Machine Learning Engineer

samhithanuka51@gmail.com +1 812-361-5059 SamhithaNuka github.com/snuka75 Portfolio

PROFESSIONAL SUMMARY

Data Scientist with 2+ years of experience applying machine learning and statistical analysis to real-world challenges in fraud detection, cybersecurity, and NLP. Built pipelines handling 10K+ logs/min, automated ETL workflows for 7M+ records, and deployed GenAI apps with 30–50ms latency. Certified in AWS ML and Databricks DE, with hands-on expertise in Kafka, Spark, and Databricks. Achieved 84%+ model precision and created dashboards featuring 20+ visualizations to drive data-informed decisions.

EDUCATION

MS in Data Science, *Indiana University Bloomington* Aug 2023 – May 2025 | Bloomington, United States

SKILLS

Data Engineering: ETL/ELT (Databricks, Spark, Informatica), Data Warehousing (Snowflake, Redshift, BigQuery), Stream & Batch Processing (Kafka, Hadoop), NoSQL (MongoDB, Cassandra), Orchestration (Airflow)

Machine Learning: Supervised and Unsupervised Learning Algorithms,Long Short term Memory(LSTM), NLP, CNN, LLM(GeminiPro,Gemma,Groq), Transformers, Random forest, K-mean Clustering, SVM.

Generative AI & Agentic Systems: LLMs (Gemini, Gemma, Groq), Prompt Engineering, RAG, FAISS, LangChain, Agentic AI Workflows

Data Analysis & Visualization: Excel (Pivot Tables, Power Query, VLOOKUP), Tableau, Power BI, Looker Studio, Pandas, NumPy, Matplotlib, Seaborn, SAS

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

WORK EXPERIENCE

Indiana University Bloomington, *Graduate Research Assistant* Jan 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.
- Engineered 15+ statistical features (e.g., inter-event intervals, rolling variance, 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* Aug 2021 – Jul 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 **<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

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

- Built **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.**

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%**.

YOLOv7-Based Object Detection for Face Mask Compliance, *YOLOv7 | Computer Vision | OpenCV | CNN | Tensorflow*

- Fine-tuned YOLOv7 on a custom Kaggle face mask dataset annotated using CVAT, **achieving ~84% mAP@0.5** and strong real-world detection accuracy.
- Optimized training pipeline with **data augmentation, hyperparameter tuning**, and performance tracking using TensorBoard.

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