Samhitha Nuka

samhithanuka51@gmail.com | +1 812-361-5059 | SamhithaNuka | github.com/snuka75 | Bloomington, United States

EDUCATION

MS in Data Science, *Indiana University Bloomington* Aug 2023 – present | Bloomington, United States
GPA: 3.6/4

Bachelor's of Technology in Electronics and Communication Engineering, Aug 2017 – Jun 2021 | Hyderabad, India
VNR Vignana Jyothi Institute of Engineering and Technology
GPA - 9.2/10

SKILLS

Data Engineering: Databricks, ETL, ELT,MangoDB,Cassandra,Apache Spark, Hadoop, Kafka, **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., **Machine Learning Frameworks:** TensorFlow, PyTorch, **Data Analysis:** Excel(Pivot Tables,Power Query,VLOOKUP), SQL, Tableau , Power Bi, **Cloud Services:** Azure Databricks, AWS(SageMaker,Glue,S3,Athena,Redshift,Lambda,MLOps), Snowflake Data Cloud, GCP, BigQuery, LookerStudio, **Programming Languages:** Core Java, HTML, CSS, JavaScript, PHP, Python, R, **Framework:** Spring, Spring Boot, NodeJs, **Web Services:** Rest API, SOAP, **OpenText Content Server 20.4**, **Social Listening:** Text Mining, Trend Detection, Custom Topic Tagging

WORK EXPERIENCE

Indiana University Bloomington, Jacobs school of Music, *Research Assistant* Jan 2025 – present | Bloomington, United States
- Programmed custom Python scripts to prepare 10,000+ beat timestamps for marimba classification.
- Engineered time-series features and trained Random Forest, SVM, and Logistic Regression models, achieving 92% classification accuracy after hyperparameter tuning with GridSearchCV and 5-fold cross-validation.
- Modeled tempo fluctuations across eight analyzed songs using linear regression, pinpointing timing discrepancies and informing targeted adjustments to enhance rhythmic precision for improved musical performances.
- Visualized model predictions and beat distributions using Matplotlib and Seaborn, clarifying regression trends and enhancing stakeholder understanding of key analytical insights across eight analyzed songs.

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 that automated data extraction and analysis, becoming the standard for future integrations.
- Enhanced legacy systems to support ML workflows, improving processing efficiency by 30% and boosting search index speed by 40%.
- Orchestrated serverless inference endpoints leveraging API Gateway and Lambda, maintaining less than 100ms latency and 99.9% uptime while processing over 500 concurrent requests daily, ensuring seamless ML pipeline operation.
- Set up real-time monitoring with Amazon CloudWatch and streamlined deployment processes through AWS CodePipeline, accelerating model release cycles.
- Performed advanced SQL diagnostics on Amazon RDS to resolve data consistency and pipeline issues, ensuring high data availability.

ACADEMIC PROJECTS

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.

Blog Generation using Amazon Bedrock and LLaMA, *Amazon Bedrock | LLaMA 3 70B Instruct | AWS Lambda| API Gateway | Amazon S3*
- Built a serverless blog generation application using AWS Lambda, API Gateway, and Amazon Bedrock (LLaMA 3 70B Instruct), achieving <200ms response latency and eliminating infrastructure management overhead.
- Integrated Postman for testing and Amazon S3 for logging, enabling full traceability of prompt/response data and ensuring scalable, production-ready LLM deployment.

End-to-End AI Application Suite for Multimodal Interaction, *Streamlit | Gemini Pro/Flash | FAISS | Groq | LLM*
- Developed a suite of 4 AI-powered web applications using Streamlit, enabling natural language Q&A, PDF analysis, image interpretation, and no-code SQL query generation with Gemini Pro/Flash and SQLite.
- Integrated Groq and Gemma for high-speed document processing with response times as low as 30–50 milliseconds, and leveraged FAISS to enable semantic search across 30+ documents with sub-second retrieval times.

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%.

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%.

NYC Crime Analytics Dashboard,Cloud-Based Public Data Pipeline, *BigQuery | SQL | Looker Studio | ETL | Feature Engineering | GCP*
- Developed an end-to-end ETL pipeline to ingest and transform 7M+ crime records, applying 12+ SQL-based feature engineering steps for time, location, and demographic analysis.
- Created a 3-page interactive dashboard in Looker Studio with 20+ visualizations, enabling real-time exploration of crime trends across boroughs and time segments.

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