AMAN WAOO

Data Scientist | Predictive Analytics | Statistical Modeling | Critical Problem Solving
Python | SQL | R | Dataiku | Airflow | AWS | Power BI | Tableau | SAS | Gen AI | Git | MS Excel
(+1) 919-985-9088 | Raleigh, NC | amanwaoo98@gmail.com | Portfolio | linkedin.com/in/aman-waoo | github.com/amanwaoo

PROFESSIONAL EXPERIENCE

North Carolina State University (Sozzani Lab), Raleigh, North Carolina — September 2023 – Present
Machine Learning Researcher

- **Statistical Modeling:** Developed custom CNN-Attention-BiLSTM (**TensorFlow**) classification model to categorize protein sequences, achieving F1 score of **96%.** Incorporated **SHAP** for network explainability (*Nature Journal).*
- **Experimental Design:** Collaborated with cross-functional research teams to develop biologically informed **Bayesian** probabilistic networks (**PyTorch**), successfully predicting agronomic traits for Soybean plant with **88%** accuracy.
- **Pipeline Automation:** Engineered and automated the data preprocessing workflows on the High-Performance Computing (**HPC**) cluster using **Bash scripting**, reducing manual effort by **80%** and improving process efficiency by **25%**.
- **Data Analytics:** Conducted Exploratory Data Analysis (**EDA**) on over 50,000 genomic data samples of Arabidopsis plant, leveraging **Apache Airflow** for its capabilities in data orchestration and **ETL** framework.

Schlumberger (SLB), Houston, Texas — May 2023 – August 2023
Data Scientist

- **Time-Series Forecasting:** Built **classical ML models** (xgboost, random forest, GBM) to predict sensor Remaining Useful Life (**RUL**), reducing oil site recall rate and achieving projected **cost savings exceeding $1M.**
- **Data Reporting:** Created **SQL** reports and interactive **Power BI** dashboards from over 1,500 electronic channels, providing senior management with actionable insights into key maintenance scheduling metrics.
- **Deep Learning Models:** Experimented with **Transformers-based** deep learning architectures (Python) for seamless integration into the Health Analyzer tool, deploying models using **Dataiku** packages.
- **Data Quality Assurance:** Mitigated data imbalance issues by using sampling techniques and generating realistic synthetic datasets. Utilized advanced statistical methods (histograms, probability mass function) to drive **data-informed decisions.**

Larsen & Toubro (L&T), Gujarat, India — August 2020 – June 2022
Data Analyst - Digital Initiatives Group

- **Design Optimization:** Led relay coordination studies for thermal power plants, using tools in **R** (ggplot, tidyverse), **HiveQL** (Hadoop), and PLCs, achieving a **30%** reduction in operational downtime. Proposed data-driven strategies resulted in cost savings exceeding **$0.5M** across **6 projects.**
- **Preventive Maintenance:** Developed **regression models** (Python) to forecast relay time-to-trip, enabling proactive sensor replacement. Enhanced plant protection reliability by reducing unplanned outages.
- **Data Visualization:** Developed a dynamic **Power BI** dashboard to track Key Performance Indicators (**KPIs**) in real-time, improving stakeholder engagement and reducing report generation time by **25%.**

ACADEMIC PROJECTS

- **Large Language Models – ChatGPT Clone for Question-Answering Application** | OpenAI, MongoDB
  - Developed a question-answer **RAG** system leveraging Large Language Models (**LLMs**) to retrieve and generate accurate responses from user-specific data, achieving **15%** improvement in response accuracy.
- **Deep Learning – Deepfake Images Detection Algorithm (GitHub) |** PyTorch, OpenCV
  - Generated **120,000** fake images from CelebA dataset for different **GANs**, used **AWS SageMaker** and trained a combined Siamese Network for fake/real images classification, achieving F-1 score of **98%**.

EDUCATION

North Carolina State University, Raleigh, USA, Masters (Electrical Engineering), GPA 4.0 — August 2022 – May 2024
Coursework: Machine Learning, Data Analytics, Data Science, Deep Learning, Signal Processing, Computer Vision

National Institute of Technology, Bhopal, India, Bachelors (Electrical Engineering), Top 10% — July 2016 – June 2020

PUBLICATION

- Morffy, N. *et al.* "Identification of plant transcriptional activation domains." *Nature (2024).*

AWARDS AND CERTIFICATIONS

- **2nd place** in the Machine Learning track at the annual N.C. Plant Sciences **Hackathon** to predict crop yield, powered by SAS.
- Dataiku DSS – "Machine Learning Practitioner", "Core Designer", NVIDIA – "Deep Learning Fundamentals", "Image Segmentation Techniques", "Time Series Data Modeling with RNN", DeepLearning.AI – "Generative AI with LLMs".