ARSALAN ANWAR

New York, NY | 9295781098 | ax2134@nyu.edu | linkedin.com/in/arsalan-anwar-ai/ | github.com/arsalananwar11 |
arsalananwar.pythonanywhere.com

EDUCATION

New York University *New York, NY*
Master of Science in Computer Science *Sep 2023 - May 2025*

Visvesvaraya Technological University *Bangalore, IN*
Bachelor of Engineering in Computer Science and Engineering *Aug 2016 - Aug 2020*

SKILLS

Languages: Python, C#, Java, C, C++, SQL, PySpark, SparkSQL
Frameworks/Libraries: TensorFlow, PyTorch, Keras, Flask, OpenCV, Docker, Git, MLFlow, REST API
Data Visualisation Tools & Cloud Technologies: Power BI, Tableau, Azure, AWS, Snowflake
Other Skills: Data Warehousing, DevOps, Postman, MLOps, Explainable AI, Big Data, Hypothesis Testing, Agile Development

WORK EXPERIENCE

New York University **New York, NY**
Data Scientist (Research Assistant) *Jan 2024 - Present*
- Spearheaded the development of NYU Mate, a course assistant built over the Llama 2 Large Language Model (LLM), leading to an average 21% improvement in student grades among a cohort of 100 students.
- Implemented a custom response pipeline using **Retrieval Augmented Generation (RAG)** with **LangChain** and **Chroma** vector database for faster data retrieval and created a user-friendly Flask interface deployed on Azure, boosting student satisfaction by 18%.

Medidata Solutions **New York, NY**
Data Science Intern *May 2024 - Aug 2024*
- Streamlined consent form completion for clinical trials by integrating **advanced NLP techniques** and **LLMs**, reducing initial setup time by 45% and setup costs by 30%, while enhancing participant onboarding efficiency.
- Implemented **AWS Lambda functions** to extract metadata, identify required input fields, and send the consent form as a **DocuSign envelope** to the respective signer, reducing document annotation time from 15 minutes to **under 2 minutes** per document.
- Utilized **Claude AI Sonnet** using **AWS Bedrock** to summarize key consent information and explain medical jargon, decreasing participant dropout rate by 15%.

Course5 Intelligence **Bangalore, IN**
Data Scientist *Feb 2023 - Aug 2023*
- Utilized the **Apriori** algorithm for **purchase analysis** to discover key product associations/patterns for **upselling** and **cross-selling**, enhancing the customer experience and potentially leading to a 15% increase in average revenue per customer.
- Developed an ensemble of **Customer Lifetime Value (CLV)** and **churn prediction models** using agile methodology for data-driven **customer segmentation** & **risk profiling**, leading to the identification of top 5% high-value, high-risk customers with 94% accuracy.

West Pharmaceutical Services **Bangalore, IN**
Data Scientist *Aug 2022 - Feb 2023*
- Designed a robust & scalable architecture to identify & classify 15 defect categories on the rubber stoppers using the **XceptionNet deep learning model** achieving an accuracy of ~92% outperforming legacy manual classification by ~37%.
- Deployed the defect detection model in production using **Docker**, and **Azure**, and created **Power BI** reports to visually represent and communicate daily classification results and insights to stakeholders, improving the identification of production line issues by 37%.

Associate Data Scientist *July 2020 - July 2022*
- Designed and set up a document parsing system using **Natural Language Processing (NLP)** techniques such as **transformers** like **BERT, Albert,** and **Roberta** with cased and uncased language which resulted in 23% better mapping of keywords with the West blogs.
- Set up **ETL data pipelines** using **Azure Data Factory** for data orchestration where documents from 6+ sources were ingested, transformed, and loaded onto a single data lake.

Graduate Software Trainee *Jan 2020 - July 2020*
- Trained and deployed Particulate classification models for structured and unstructured data using **Azure ML** and **Azure Function Apps**. Achieved an overall accuracy of ~97% through the modeling, optimization, and evaluation process.
- Replaced the manual classification process used by the Lab Analysis team with the new models, resulting in a reduction of 80% in the time required for particulate classification tasks.

ACADEMIC PROJECTS

Safeguarding NYC: Analyzing Crime Patterns using Big Data | *Python, SQL, Spark, Snowflake, Data Analysis, Big Data, Power BI*
- Projected a 30% enhancement in NYPD's crime prevention efficiency through big data analysis of 220k+ crime records using advanced ELT pipelines and Spark.

Aqua Analytica: Water Consumption Analytics System for Bengaluru | *Flask, Time Series, Demand Forecasting, Machine Learning*
- Achieved 25-30% water savings (~2.1 billion liters) in Bengaluru through a smart analytics system and innovative water management strategies.

LEADERSHIP, AWARDS AND RECOGNITIONS
- Delivered a seminar on "Getting Started with Image Classification Using Azure ML" at the **Microsoft User Group Hyderabad**, providing insights into the fundamentals of image classification and Azure Machine Learning
- **Top 3** under Computer Science Category **all over India** for Innovative Student Projects Award organized by the prestigious Indian National Academy of Engineering (INAE) among thousands of teams taking part