Mohit Kundu

United States | github.com/Mohit-Kundu | linkedin.com/in/mohit-kundu-cs

https://medium.com/@moh8.kundu | https://mohit-kundu.github.io/

Education

MS in Computer Science, *Indiana University* 08/2022 – 05/2024
Data Structures, Algorithms, Machine Learning, Data Mining, Data Science, Database Management, Software Engineering, Information Visualization, Data Modelling Bloomington, Indiana

Professional Experience

Data Scientist, *Project 990* 08/2024 – present
- Designing **ETL pipelines** for **data aggregation** using **SQL** and **Pandas** to extract, transform, and load **over 2 million rows** from the IRS API into **databases**.
- Applying **causal inference** and **testing for multicollinearity** to identify key factors driving nonprofit outcomes, supporting data-driven recommendations.
- Building **ensemble classification models** and **fine-tuning LLMs (Llama-2, Mistral 7b)** to categorize nonprofit programs with **92% accuracy**.

Chicago, United States

Research Assistant - Data Science, *Kelley School of Business - Indiana University* 01/2024 – 04/2024
- Developed a **probabilistic recommendation system** to predict consumer behavior.
- **Scraped** and **analyzed retail data** to assess pricing and promotions impact on sales.
- Applied **clustering** methods and **visualized** insights to inform data-driven decisions.

Bloomington, United States

Machine Learning Intern, *Hearthealth Technologies* 08/2021 – 07/2022
- Developed **deep learning** functions with **OpenCV**, **Scikit-learn** and **TensorFlow** to identify heart disease in heart scans.
- Achieving **95% precision** and a **70% reduction in diagnosis time**.
- Led weekly **Scrum meetings with stakeholders**.

Bangalore, India

Data Science Intern, *Sparkflows.io* 02/2021 – 05/2021
- Developed **ML pipelines for Big Data** systems using **Scikit-learn**, **Hadoop** and **Apache Spark**, **improving accuracy by 25%**.
- **Optimized preprocessing** and **hyperparameters** to **reduce prediction errors** by **15%**.
- Built **CI/CD pipelines** and **Flask APIs** for seamless model integration.

Santa Clara, United States

Skills

Programming Languages
Python, R, JavaScript, C++

Frameworks
Tensorflow, Pandas, SQL, AWS, Keras, Scikit-Learn, PySpark, Pytorch, Matplotlib, MLFlow, Airflow, Kafka

Projects

Research Paper Query System with RAG Architecture
- Built a **RAG** with **Llama-2**, **Milvus DB, CohereEmbeddings** achieving **90%+ accuracy**, evaluated via **ARES**.
- **Improved accuracy by 25%** with **semantic chunking**, while **reducing query retrieval time by 30%**.
- Developed a **Streamlit** interface and containerized the system with **Docker** for consistent, scalable deployment.

Automated Stock Forecasting Pipeline with AWS and Airflow
- **Scraped stock data** using Yahoo Finance **API** with **Apache Airflow**, storing it in **S3** buckets.
- Built an **ETL pipeline** with **AWS Glue** and trained a **CNN-LSTM** model on **SageMaker**, achieving **93% forecasting accuracy**.
- Streamlined decision-making with **QuickSight visualizations**.

US Home Ownership Constraint Analysis (Indiana Business Research Center)
- Developed a **clustering** and **dimensionality reduction pipeline** to identify key constraints.
- Built intuitive **visualizations** using **Tableau** and **Power BI**, enabling **data-driven decision-making**.

Event Management System using MERN stack
- Led development of **RESTful APIs** for event management using **Node.js**, **Express.js**, and **React.js**.
- **Improved search time by 17%** with **MongoDB** and **ElasticSearch**.
- Ensured reliability and security through comprehensive **API testing** with **Postman** and **ApiDog**.