Anusree Mondal Rakhi

ar4636@columbia.edu | +1 (631) 720-8838 | Legal Permanent Resident

PROFESSIONAL SUMMARY

Data Scientist and Columbia University graduate with 2 years of experience delivering end-to-end machine learning (ML), large language models (LLMs), and retrieval augmented generation (RAG) solutions. Recently, I dived deeper into AI agents and completed a Google course

EDUCATION

Columbia University New York, NY
Master of Science in Data Science **Sep 2023 - Feb 2025**
Relevant Coursework: Exploratory Data Analysis & Visualization, Statistics, Machine Learning, Computer Vision, DL for NLP

SRM Institute of Science and Technology Chennai, India
Bachelor of Technology in Computer Science and Engineering **Jul 2018 - May 2022**
Relevant Coursework: Data Mining and Analysis, Database Management Systems, Design and Analysis of Algorithms

TECHNICAL SKILLS

Python, R, SQL | NumPy, pandas, SciPy, scikit-learn | TensorFlow, Keras, PyTorch | spaCy, NLTK, OpenCV | LangChain, Hugging Face, FAISS, Embeddings, LLMs, RAG, AI Agents | Regression, Classification | AWS (SageMaker, S3, Redshift, MLflow), GCP | MLOps (automation, retraining, monitoring) | Web Scraping (requests, BeautifulSoup, Selenium) | Statistical Modeling, Hypothesis Testing, A/B Testing, Experimental Design, Evaluation Metrics, Data Mining | Tableau, Power BI, Matplotlib, seaborn, ggplot2 | Data Quality and Validation | GitHub, Excel

WORK EXPERIENCE

Mastercard | *Applied AI Scientist* | *New York, NY* **Sep 2024 - Dec 2024**

- Delivered a web-scraping and text-only ingestion pipeline for **10,000 URLs** using BeautifulSoup batching with Selenium fallback
- Optimized multilingual translation by benchmarking **6 translators** across 35 languages and productionizing the best performing translator, reducing end-to-end processing from **4 weeks to 3 hours**
- Enabled cost-efficient website categorization using LLMs; evaluated keywords with synonym-aware Precision@3 against a human-annotated dataset, lowering run costs from **$46 to $2.30** per 10,000 URLs and achieving **~16 to 20 min** runtime with concurrency

Columbia Climate School | *Data Scientist* | *New York, NY* **Sep 2024 - Dec 2024**

- Developed a **Streamlit-based RAG application** for environmental researchers and climate scientists to ingest multiple articles and reports and answer queries on scope of work, constraints, eligibility, and assumptions with source-cited responses
- Optimized document review by chunking and embedding documents and retrieving only the top relevant evidence per query, reducing OpenAI API token cost versus sending full reports and cutting turnaround time by **84%**
- Optimized an existing CNN model by strengthening the convolution operation with stacked kernels, BatchNorm, Dropout, and EarlyStopping; increased accuracy by **8%** and cut training time by about **15%** without new data collection

BlocPower | *Data Scientist* | *Brooklyn, NY* **May 2024 - Aug 2024**

- Led data validation on **120M+** records in Amazon Redshift (SQL), diminished data inconsistencies by **34%** and improved data quality
- Developed and productionized supervised ML models (S3, Amazon SageMaker) to predict key building energy characteristics supporting decarbonization efforts
- Integrated model lifecycle into a production **MLOps** pipeline (MLflow) with a **CI/CD** workflow to automate retraining, validation, and deployment; enabled assessments that helped clients reduce energy waste by up to 80%
- Created performance monitoring reports for **cross-functional teams** (Engineering, Ops) providing actionable insights into critical business metrics

Columbia Engineering | *Machine Learning Research Assistant* | *New York, NY* **Sep 2023 - Apr 2024**

- Reduced Teaching Assistants' workload by **50%** in high-enrollment courses by answering students' queries through an LLM assistant
- Enabled fast, accurate semantic retrieval using a LangChain LLM + RAG pipeline with embeddings and a vector database over an Excel knowledge base
- Improved answer reliability with Google PaLM using a strict context-only prompt and "I don't know" fallback, deployed via Streamlit with an admin re-index workflow
- Derived insights from **100,000** public health records across **30 years** by building visualizations, linking demographic and social factors to lifestyle changes, and presenting findings through **Tableau** dashboards

Gonit Shikkha Kendro | *Data Science Instructor* | *Dhaka, Bangladesh* **Jun 2022 - Aug 2023**

- Taught data preparation for ML, DL, and NLP models using Python, SQL, NumPy, pandas, seaborn, scikit-learn, spaCy, NLTK
- Designed and led workshops on **data storytelling**, **critical thinking**, and **strategy development** for real-world data problems

SRM Institute of Science and Technology | *Deep Learning Research Assistant* | *Chennai, India* **Jun 2021 - May 2022**

- Enhanced Mask-RCNN on satellite imagery by optimizing Backbone architecture (**ResNet**) and model **hyperparameter tuning**
- Yielded **82%** object detection accuracy, demonstrating strength in small object recognition, and published results with Springer