Ritwik Mishra

Leeds, UK — mritwik11@gmail.com — 07934 547923
linkedin.com/in/ritwikm4 — github.com/ritwik4m

Professional Summary

Data Scientist with 5+ years of experience delivering cloud-based, analytical, and quality-assured data solutions. Skilled in Python, SQL, Azure, and Databricks, with practical expertise applying machine learning, NLP, and RAG techniques to extract insights from complex datasets. Adept at building reliable data pipelines, leading CI/CD adoption, and collaborating in Agile and DevOps environments to deliver scalable analytics for regulatory and financial domains.

Technical Skills

- **Programming Languages:** Python, SQL, Java
- **Data Science & Analytics:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Power BI, Tableau, NLP (text preprocessing, sentiment analysis, topic modeling), Statistics, Data Mining
- **Machine Learning:** Supervised/Unsupervised Learning, Feature Engineering, Model Evaluation, Transformers (BERT, Hugging Face), Deep Learning (PyTorch, TensorFlow), Data Preprocessing
- **Data Engineering:** Databricks, ETL validation, data pipeline automation, data quality assurance, CI/CD integration
- **Cloud Platforms:** Microsoft Azure (Data Factory, Databricks, Synapse), Azure AI Fundamentals certified, AWS
- **QA & Automation:** Test planning, API testing, automation scripting, Selenium, Playwright, Cucumber BDD, Jira, HP ALM
- **Soft Skills:** Analytical thinking, stakeholder communication, collaboration, attention to detail

Experience

Data Engineer *May 2022 – Aug 2024*
Paxcom India (A Paymentus Company), Gurgaon, India

- Led validation and monitoring of Azure Databricks-based data pipelines ensuring analytics data reliability across distributed systems.
- Automated Python and SQL workflows for ETL testing and anomaly detection, improving coverage and data integrity.
- Collaborated with data engineers to optimise ingestion and transformation pipelines via Azure Data Factory.
- Supported CI/CD adoption by integrating automated validation into DevOps workflows, reducing deployment defects by 20%.
- Designed Power BI dashboards to visualise data quality metrics and testing KPIs for leadership reporting.

Data Analyst *Nov 2018 – Apr 2022*
Accenture Solutions Pvt. Ltd., Pune, India

- Delivered backend and ETL testing within Agile sprints, ensuring accuracy and consistency of large-scale telecom CRM data.
- Developed SQL-based validation scripts to verify data integrity during database migrations and analytics transformations.
- Collaborated across business and technical teams to translate data requirements into automated test frameworks.
- Contributed to CI/CD pipeline testing and documentation for continuous deployment environments.

Education

MSc Data Science *Sept 2024 – Sept 2025 (awaiting results)*
University of Surrey, Guildford, UK

- Modules: Database Systems, Business Analytics, Data Mining, Machine Learning, NLP, Cloud Computing.
- **Dissertation:** *BioDynaMo-Based Pipeline for Neuron Morphology Optimisation* – Developed a Python-based simulation workflow integrating BioDynaMo for large-scale data analysis and model optimisation.

BEng Electronics & Telecommunication *Aug 2014 – Jun 2018*
Bhilai Institute of Technology, India

- Final-year project: Designed a wireless IoT network for polyhouse monitoring and control.

Projects

- **RAG-based Document Insight Engine** — Built a Retrieval-Augmented Generation (RAG) pipeline using LangChain, Azure Databricks, and OpenAI API to extract and summarise financial documents. Implemented embedding models for semantic search and transformer-based summarisation.
- **NER Token Classification** — Developed a BERT-based Named Entity Recognition model for text analytics, achieving an **F1-score above 0.85** on entity extraction tasks using Hugging Face Transformers.
- **House Price Prediction** — Built regression models using `Python` and `Scikit-learn` to predict housing prices across 21k+ properties. Implemented feature engineering and model tuning (Random Forest, XGBoost) to reduce RMSE and improve accuracy.
- **SaaS Data Warehouse Pipeline** — Designed an automated data warehouse pipeline integrating ingestion, transformation, and quality checks using `Python` and `SQL`. Ensured end-to-end data consistency and reliability for analytics dashboards.

Certifications

- **Databricks Lakehouse Fundamentals** – Databricks - May 2024
- **Microsoft Certified: Azure AI Fundamentals** – May 2025
- **Generative AI with Large Language Models (LLMs)** – Jun 2024
- **Postgraduate Program in Data Science and Business Analytics** – Great Lakes Institute of Management
- **Postgraduate Program in Data Science and Business Analytics** – University of Texas at Austin

Additional Information

- Eligible to work in the UK.
- Volunteering: National Service Scheme – literacy & blood donation drives.
- Volunteering: University of Surrey – community gardening and sustainability projects.
- Interests - AI for Sustainability