Sudhanshu Mukherjee

sudhanshuxmukherjee@gmail.com | Phone: (+1) 781-580-3292 | Boston, MA
https://www.linkedin.com/in/sudhanshumukherjeexx/ | https://github.com/sudhanshumukherjeexx

EDUCATION

- University of Massachusetts Dartmouth, MA, USA – Master of Data Science, Sep 2022 – May 2025
- G.H. Raisoni College of Engineering, Nagpur, India – Bachelor of Engineering - Computer Science, Aug 2015 – May 2019

TECHNICAL SKILLS

- **Programming Languages:** Python, SQL, C, Bash
- **Interests & Experience:** Predictive modelling (Random Forest, XGBoost, Clustering, etc), Fine-Tuning LLMs, Natural Language Processing, A/B testing and experimental design, Time Series forecasting (ARIMA, Prophet)
- **Frameworks:** PyTorch, Transformers, PySpark, Scikit-Learn, Pandas, Polars, NumPy, Langchain, Langgraph, MLFlow, wandb, Statmodels, Matplotlib, Streamlit, Gradio
- **Tools/Platforms:** Microsoft Power BI, AWS (Sagemaker, S3, Redshift), Microsoft Azure, Docker, Kubernetes, CI/CD, Git
- **Applied Statistics:** Probability theory, Calculus, Linear Algebra, Bayesian Forecasting
- **Certification:** AWS Machine Learning Spec.

PROFESSIONAL EXPERIENCE

Data Scientist, Community Dreams Foundation, Remote, Sep 2025 – present

- Engineered an **autonomous AI health agent** using **LangGraph** to orchestrate a multi-step workflow, automating the end-to-end process from data ingestion to personalized health recommendation generation
- Developed a comprehensive data analysis pipeline using Python and Pandas to ingest, synthesize, and analyze disparate biomedical data streams, including wearables, structured labs, genomics, and user surveys
- Designed a **rule-based algorithm** to calculate a holistic health score and integrated a **Google Gemini LLM** to translate complex data profiles into actionable, prioritized lifestyle interventions for users

Data Scientist (Summer Intern), Camping World – Good Sam, Chicago, IL, June 2023 – Aug 2023

- Engineered a **K-modes clustering** model using python using 10K+ RV inventory records, achieving 85% internal similarity, supporting segmentation for targeted advertising, marketing, and sales personalization
- Assisted in validation of scalable **ETL pipelines** for RV datasets using SQL, Snowflake, and SQL Server Management Studio (SSMS), improving data reliability and accuracy for downstream ML workflows
- Collaborated in building a real-time enterprise dashboard in Power BI with custom DAX metrics, enabling 30+ executives to monitor sales KPIs and model performance across regions

Data Scientist, Unosis IT Solutions Private Limited, India, July 2019 – May 2022

- Designed and deployed end-to-end **predictive models** (regression, classification) for **churn, purchase intent, and customer lifetime value (CLV)** on an e-commerce platform with 10M+ users, driving measurable improvements in system performance and user engagement
- Applied **BERT-based NLP** models to analyze product search queries and session text, translating technical performance metrics into improved search **relevance and user satisfaction**
- Implemented **ARIMA and Meta Prophet forecasting pipelines**, enhancing inventory planning, pricing strategies, and demand forecasting, integrating results into the enterprise data warehouse for downstream analytics
- Applied **causal inference, experimental design, hypothesis testing, and A/B testing frameworks** to evaluate model effectiveness and measure performance impact at scale
- Engineered and optimized **complex SQL queries** and ETL pipelines for large-scale data transformation, performance tuning, and real-time system monitoring
- Developed and maintained **data warehousing solutions** (dimension modelling, star/snowflake schemas) to support scalable, high-performance analytics on large datasets and streaming data
- Implemented **MLOps pipelines** using **Kubernetes, Jenkins, and AWS Lambda** for distributed model deployment, enabling serverless inference at scale and automated CI/CD workflows that improved system performance and reduced latency for production workloads
- Orchestrated **ML experimentation** and **reproducibility** using MLFlow, enabling model versioning, registry, and performance tracking across distributed systems
- Developed **Power BI dashboards** and monitoring tools to translate complex technical and performance metrics into clear insights for **engineering, product, and business stakeholders**
- Collaborated **cross-functionally** with engineering teams and leadership to translate business requirements into **actionable recommendations** for improved user experience and business outcomes

Software Engineering Intern, Cojag Smart Technology, India, June 2018 – Feb 2019

- Designed and implemented responsive, user-friendly web interfaces and UI mockups, improving site usability and user engagement. Contributed to full-stack development, integrating frontend functionality with backend databases for seamless real-time performance
- Collaborated on an IoT-based automated food ordering system using Python and MySQL, reducing manual intervention and streamlining restaurant operations from order placement to delivery tracking

LEADERSHIP EXPERIENCE

Teaching Assistant – DSC 520, University of Massachusetts Dartmouth, MA, Sep 2024 – May 2025

- **High Performance Scientific Computing:** Mentored 75+ grads with implementing projects in **PyTorch**, Python optimization, and High-Performance Computing topics like **parallel processing**, **memory hierarchies**, **Linux command line**, and multi-core systems

Data Instructor, Digital Scholarship Hub, UMass Dartmouth, Dec 2023 – Aug 2024

- Led workshops on various tech topics like Data Visualization, ArcGIS, Machine Learning, SQL, and LangChain for faculty, students, and non-technical staff
- Developed Open Access guides for complex tools such as Git and data analysis, helping students and staff use them in their work
- Collaborated with the Scholarly Communication Librarian for research, reports, troubleshooting issues, maintaining digital reports, and auditing LibGuides

Teaching Assistant – CIS 190, University of Massachusetts Dartmouth, MA, Sep 2023 – Dec 2023

- **C Programming:** Led hands-on C programming labs, teaching key concepts like data types, variables, operators, and control structures to undergraduate students. Developed topic-specific labs, explained coding tasks, and graded assignments

MASTERS THESIS, Jan 2023 – May 2025

Rapid Insight Data Engine: An open-source Python framework for analyzing Tabular Data - thesis

- **R.I.D.E.:** Built and maintain a no-code machine learning platform designed to help users implement a data analysis workflow to explore, clean, transform, and model data (CSV, Excel, and Parquet) backed by GPT-4 to analyze results and maintain memory for users to chat with data. Deployed using Docker and Kubernetes for scaling with users from my university - docs
- **RIDE-CLI:** Menu-driven CLI Python package that allows you to perform Feature Engineering on data (CSV, Excel, and Parquet), such as Data Quality checks, feature scaling and transformation, handle missing values, and perform regression and Classification through AutoML for streamlined data analysis workflows with 500+ downloads - docs

FEATURED PROJECTS

RAG System - Document Summarization and Self-Corrected Question Answering - GitHub

- Built an intelligent document analysis system with a Gradio UI, featuring a self-correcting RAG pipeline using GPT-4. The tool delivers high-accuracy document Q&A and summarization, ensuring response quality with automated scoring and a guardrail agent

Multimodal Medical Vision Question Answer with LLaVA Fine-tuning - notebook

- Fine-tuned a large vision-language model (LLaVA-1.5-7B) using Low-Rank Adaptation (LoRA) for medical visual question answering on radiology images. Implemented a custom evaluation pipeline with token-level F1 scoring, managed training workflow on Google Colab with PEFT library, and deployed model artifacts to Hugging Face Hub for reproducible research

American Express Kaggle Dataset: Insights into Fraudulent Transaction Detection - GitHub

- Built and analysed the effectiveness of machine learning algorithms such as XGBoost, CatBoost, LSTM, and logistic regression in detecting fraudulent transactions using the American Express Kaggle dataset (5.5 million data points and 192 features). CatBoost outperformed XGBoost with an accuracy of 98.1%, compared to XGBoost's accuracy of 97.7%

CIFAR100: Comparative Analysis of Deep Learning Architectures - GitHub

- Conducted comprehensive performance analysis across 14 deep learning architectures such as CNNs, EfficientNets, Vision Transformers, and Mobile Networks, implementing robust measurement frameworks and statistical comparisons to evaluate model effectiveness and efficiency metrics

CO-CURRICULAR

- **Data Carpentries Certified Instructor** – May 2025