

# Arijit Mukherjee

857-350-6469 • [arijit97@hotmail.com](mailto:arijit97@hotmail.com) • [linkedin.com/in/arijit-mukherjee](https://www.linkedin.com/in/arijit-mukherjee) • [github.com/ironbongir](https://github.com/ironbongir)

## Education

---

### Northeastern University, Boston, MA

Master of Science, Data Analytics Engineering

April 2024

- Coursework: Data mining, machine learning, Operations research, Algorithms, MLops

## Technical Skills

---

<b>Languages</b>	Python, R, MySQL, Postgres, Linux, Bash, Spark, PySpark
<b>AI/ML</b>	Pandas, Numpy, Scikit, Neural Networks, LLM, RAG, FAISS, Tensorflow
<b>Technologies</b>	Control-M, Airflow, Tableau, Git, Hadoop, REST, CI/CD
<b>Cloud</b>	AWS, S3, EMR, Athena

## Work Experience

---

### Deloitte

August 2019 - August 2022

Analyst (Analytics Engineer)

- Built and optimized ETL pipelines using SQL and Spark on AWS, improving data processing efficiency by 50% and ensuring reliable data workflows.
- Presented complex data insights to non-technical stakeholders translating analytics into actionable recommendations that reduced operational bottlenecks by 45%.
- Spearheaded the design and optimization of complex data pipelines across 4 critical modules, ensuring data transformation, modeling, and integration while managing dependencies to drive informed business decisions.
- Supported and Monitored data pipelines, including on-call responsibilities to ensure data flow continuity and timely issue resolution.
- Collaborated cross-functionally to design and implement data experiments for clients, driving business impact with insights and solutions.
- Architected automated solution for non-standard directory removal in AWS S3, slashing manual work by 60% and enhancing data management processes.

## Projects

---

### DunderFetch

October 2024 - October 2024

- Developed a Q/A bot using Retrieval-Augmented Generation (RAG) for efficient semantic search and content retrieval from The Office script.
- Leveraged SentenceTransformer (all-MiniLM-L6-v2) and FAISS to improve content matching accuracy by 20%, facilitating fast and precise similarity searches.

### FairLens CreditAI

January 2024 - April 2024

- Designed an end-to-end MLOps pipeline to predict credit defaulters, incorporating AI ethics to reduce bias and false positives by 30%.
- Conducted A/B testing and causal inference analysis to evaluate model improvements, driving better decision-making in fair lending practices.

### Predictive Analysis of Diabetes Health Indicators

October 2023 - December 2023

- Analyzed 70,000+ CDC survey responses to identify key diabetes risk factors, achieving an 85% accuracy rate with a custom logistic regression model.
- Applied statistical methods, including elastic nets, to enhance prediction accuracy by 20% over baseline models, demonstrating strong data manipulation skills.

### WeatherWise: Automated Weather Data ETL Pipeline on AWS

July 2023 - July 2023

- Devised and executed an ETL data pipeline on the AWS cloud platform, orchestrating the extraction of real-time weather data from the Open Weather Map API with precision.
- Leveraged Apache Airflow to automate the data workflow, ensuring data transformation and loading into an S3 bucket.

### Predicting Smoking Status from Medical Data

January 2023 - April 2023

- Leveraged EDA and machine learning techniques, forecasting a smoker with 92% accuracy through model optimization.