# BINDUJA MALEMPATI

NY, USA | GitHub | LinkedIn | malempatibinduja@gmail.com | 716-380-4028

## SUMMARY

Data Scientist is skilled in designing and operationalizing machine-learning solutions that drive last-mile logistics, prescriptive analytics, and customer experience improvements. Master's-level expertise in statistics and AI coupled with two years of experience deploying predictive, ensemble, and NLP models at scale. Proven record of building ETL pipelines, MLflow-tracked experiments, and continuous-integration/continuous-delivery workflows on AWS, GCP, and Docker/Kubernetes while translating complex business problems into production-grade data assets with Python, Spark, and SQL. Conducted peer code reviews in GitHub, enforced branch-protection rules, and automated linting and unit tests with GitHub Actions to uphold code quality across a distributed team. Experienced in communicating insights to non-technical stakeholders and executive leadership.

## SKILLS

Languages & Frameworks: Python, SQL, R, Scala, Scikit-learn, XGBoost, TensorFlow, PyTorch, StatsModels
Data & MLOps: Spark (PySpark, MLlib), Databricks, MLflow, Airflow, Kafka, Snowflake, BigQuery, AWS S3/Lambda, GCP Cloud Functions
ETL & Databases: PostgreSQL, MySQL, SAP HANA, dbt, Pandas, NumPy, Feature Engineering, Data Quality Checks, Parquet, Delta Lake, Feature Store
DevOps & MLOps: Docker, Kubernetes, GitHub Actions, Jenkins, continuous integration and continuous delivery (CI/CD), Terraform.
Visualization: Power BI, Tableau, Streamlit, Matplotlib, Seaborn.
Optimization & Operations Research: Gurobi, CPLEX, mixed-integer programming, constraint optimization.

## CERTIFICATIONS

AWS Certified Machine Learning – Specialty
Microsoft Certified: Azure Data Scientist Associate
Databricks Lakehouse Fundamentals

## PROFESSIONAL EXPERIENCE

**University at Buffalo**
**Data Science Analyst** **Mar 2024 – May 2025 (Buffalo, USA)**

- Engineered an ETL pipeline (Airflow DAG to Spark to Snowflake) that ingests 25 million rows per term, ensuring data-quality monitoring and schema-drift alerts.
- Developed ensemble time-series models (ARIMA and SVR) that feed daily route optimization, ETA prediction, and capacity-planning workflows for last-mile operations; registered each version in the MLflow model registry and automated retraining through continuous-integration and continuous-delivery pipelines.
- Ran cluster-scale Spark jobs on Databricks to process twenty-five million records per term, writing Delta Lake tables that feed real-time staffing and inventory forecasts.
- Packaged the inference service in Docker and deployed it to Google Kubernetes Engine, meeting a p-ninety-five latency below one hundred fifty milliseconds.
- Embedded column-level data-quality monitors and automated schema-drift alerts within the Airflow pipeline, reducing downstream dashboard errors by eighty percent and guaranteeing reliable daily reporting.

**Incture Technologies**
**Data Scientist (ML/NLP Pipelines)** **Jan 2023 – Dec 2023 (Bengaluru, IN)**

- Designed an NLP pipeline with spaCy and BERT embeddings to process daily customer feedback and surface key complaint themes feeding a **prescriptive analytics dashboard** that recommends churn-mitigation offers.
- Trained logistic-regression and support-vector-machine churn models that lifted area-under-curve to eighty-six percent.
- Deployed versioned REST services on SAP Business Technology Platform using Docker containers and MLflow.
- Produced interactive Power BI dashboards that exposed churn risk and sentiment signals to marketing and product leaders
- Authored a reusable Python package for text preprocessing that three internal teams adopted.

**Phoenix Global**
**Data Science Intern (Recommender Systems & LLM Integration)** **Mar 2022 – Aug 2022 (Hyderabad, IN)**

- Built recommender systems with collaborative filtering and content-based similarity to personalize movie suggestions for two million user ratings.
- Performed user segmentation with PySpark Gaussian Mixture Models and leveraged Gurobi mixed-integer programming to fine-tune segment-specific recommendations, boosting campaign conversion by nine percent.
- Integrated GPT-three intent recognition to route queries in customer chat and cut average support handling time by twenty-two percent.
- Designed and executed A/B tests, analyzed uplift with Bayesian methods, and presented findings to product stakeholders.

## PROJECT EXPERIENCE

**Bookstore Dashboard (Python, Streamlit, PostgreSQL, Google Data Studio)**

- Re-engineered the denormalized bookstore dataset by designing a BCNF schema in PostgreSQL and writing migration scripts with indexed joins that accelerated query performance.
- Authored parameterized SQL views and Streamlit code to render interactive sales and inventory dashboards, lowering average query latency to one point eight seconds and cutting dashboard load time by forty percent.
- Deployed the dashboard to Google Data Studio for business stakeholders, enabling daily author and genre reporting without manual spreadsheets.

**NutriWise – Nutritional Recommendation System (Python, Machine Learning, Streamlit)**

- Collected user demographic and health data, engineered nutritional features, and trained a gradient-boosted tree model to deliver personalized meal plans.
- Built a Streamlit front end with input validation and session management that raised weekly user engagement by twenty-five percent.
- Set up cloud storage and a CI pipeline that continually retrain the model as new dietary data arrives, keeping recommendations up to date.

**Prompt Health Checker (Python, NLP, Hugging Face Transformers)**

- Created an NLP scoring engine using Sentence-BERT embeddings and XGBoost to rate large language model prompts for coherence and safety.
- Implemented Spark batch jobs that evaluate one hundred thousand prompts per hour and log experiment metrics in MLflow for version tracking.
- Published a REST endpoint that delivers instant prompt feedback and reduced manual editorial review time by seventy percent.

**AI Agents (Python, Multi-Agent Systems)**

- Designed a communication protocol and task-allocation algorithm that synchronizes vision, planning, and execution agents on shared objectives.
- Implemented the Python framework with a message queue and reinforcement signals, improving cooperative task completion rates by eighteen percent in simulation.
- Wrote API documentation and example notebooks that let new researchers add custom agents within one day.

## EDUCATION

**MASTER OF SCIENCE IN DATA SCIENCE** **Jan 2024 – May 2025**

State University Of New York At Buffalo

- **Relevant Coursework:** Statistical Data Mining, Advanced Statistical Data Mining, Probability Theory and Statistics Using R, Python for Data Scientists, Data Model Query Language (SQL-focused), Data Intensive Computing, Advanced Machine Learning, Mathematical & DS Core (linear algebra, vector calculus, hypothesis testing, statistical inference, ensemble modeling, prescriptive analytics)