Jyothismaria Joseph

jyothismaria2002@gmail.com • linkedin.com/in/jyothismaria • github.com/Jyothismaria • (914) 619–3612 • Yonkers, NY

Data Science graduate with 2+ years of research experience building machine learning models, NLP pipelines, and data workflows in Python and R — with a strong foundation in statistical methods, reproducible research practices, and communicating findings to both technical and non-technical collaborators.

Education

University at Buffalo Buffalo, NY

M.S. Engineering Science – Data Science (GPA: 3.9/4.0) Aug 2024 – Feb 2026

B.S. Computer Science (GPA: 3.73/4.0) Aug 2020 – May 2024

Relevant Coursework: Machine Learning, Statistical Learning, Predictive Analytics, Data Modeling & Query Languages, Data Warehousing

Technical Skills

- **Languages & Libraries:** Python (NumPy, pandas, scikit-learn, XGBoost), R, SQL (PostgreSQL, Oracle)
- **Machine Learning & AI:** NLP (BERT, Transformers), predictive modeling, classification, clustering, regression, computer vision (CNNs, GANs), feature engineering, model evaluation (precision, recall, F1, ROC-AUC)
- **Frameworks & Tools:** scikit-learn, XGBoost, Transformers (HuggingFace), MLflow, Git/GitHub, Docker
- **Data Engineering:** ETL pipelines, data preprocessing, data validation, reproducible workflow design, Apache Hop
- **Visualization:** Tableau, Power BI, Plotly Dash, Streamlit, research dashboards

Experience

Extern Remote

AI Extern, Wayfair Mar 2026 – Present

- Built AI agents in n8n to automate trend detection and competitive tracking across the home goods sector, consolidating outputs from 5+ sources into a live dashboard — cutting trend detection time by 60%.
- Designed and documented end-to-end data pipelines for extracting, transforming, and loading multi-source data, ensuring consistency and reliability across automated reporting workflows.

Center for Computational Research, University at Buffalo Buffalo, NY

Senior Research Aide – Data Science & Analytics Feb 2025 – Nov 2025

- Built and evaluated machine learning models in Python on 100K+ record datasets to analyze system utilization patterns, improving predictive accuracy from 14% to 62% through iterative model development and cross-validation.
- Developed Python-based data pipelines for automated preprocessing and feature engineering, writing clear documentation to ensure workflows were reproducible and transferable across the research team.
- Worked closely with stakeholders to understand their analytical questions, then presented findings and model trade-offs in ways that were accessible to non-technical audiences.
- Built and maintained 5+ dashboards and recurring reports to track KPIs and system quality metrics, supporting ongoing research monitoring and team decision-making.

Summer Research Intern – Data Analysis May 2023 – Aug 2023

- Ran statistical time-series analysis of large-scale system usage logs in R to pinpoint computational bottlenecks, then reengineered the affected components in C++, reducing runtime from 40 minutes to 5 seconds.
- Published and presented the research at the PEARC Conference, where it was accepted as a peer-reviewed poster.

University at Buffalo Buffalo, NY

Teaching Assistant – Statistical Learning II Aug 2025 – Sep 2025

- Supported a graduate-level course in Bayesian statistical learning, holding weekly office hours and helping students implement clustering, regression, and Gaussian process models in R.

Project Experience

Text Sentiment & Stress Detection *Python, NLP, BERT, Transformers, scikit-learn*

- Built and compared three model architectures — RNN, Transformer, and fine-tuned BERT — for classifying stress and emotional state from Reddit and Twitter text, with BERT achieving F1 of 0.88 and ROC-AUC of 0.94.
- Used confusion matrices to systematically analyze where each model failed, using those findings to guide final model selection and improve performance on imbalanced class distributions.

Manufacturing Defect Detection Using GANs *Python, CNNs, GANs, Computer Vision*

- Trained GAN models to generate synthetic defect images, then integrated them into a CNN training pipeline to address class imbalance — improving defect recall by 22% and reducing false negatives on real manufacturing data.
- Evaluated all models using precision, recall, F1, and confusion matrices, documenting the full experimental pipeline to make results reproducible and findings easy to build on.