Lavya Midha

lavyamidha015@gmail.com | 413-800-1278
GitHub | LinkedIn | Portfolio

Education

New York University, New York, NY Expected May 2026
M.S. in Data Science (Industry Concentration) ***GPA: 3.72/4***
Relevant Coursework: Probability, Statistics, Programming, Text as Data, Big Data, Machine Learning, Practical Training for Data Science.

Boston University, Boston, MA September 2021 - May 2024
B.S. in Data Science, Minor: Statistics ***GPA: 3.75/4***
Relevant Coursework: Data Visualization, Machine Learning and AI, Algorithms, Statistical Modeling, Probability, Big Data, Data Mechanics, Optimization, Regression Analysis, NLP, Inference, Analysis of Variance.
Leadership and Awards: Undergraduate Research Grant, Dean's List, Secretary – BU Data Science Association, International Peer Mentor, Peer Tutor, Admissions Ambassador.
Recognition: Featured in BU Spark! Demo Day for civic tech innovation, recognized for public impact work in data science (Link)

Work Experience

ReferU.AI, Lewes, Delaware May 2025 – Present
Machine Learning Engineer Intern

- Built a production ML pipeline for Google Ads keyword ranking, using LightGBM and legal-domain embeddings to surface high-ROI terms in real time; deployed across 50 states, boosting efficiency by 20%+.
- Designed a semantic retrieval system for court opinions, integrating LangChain + OpenAI to deliver vector-based, citation-aware search; optimized for interpretability, speed, and legal relevance.
- Engineered a jurisdiction-aware campaign graph, linking structured court dockets with demographic features to automate ad targeting.
- Developed Looker Studio dashboards for real-time KPI tracking, aligning live Google Ads performance, web traffic, and engagement signals, reduced campaign tuning latency and improved exec decision-making agility.
- Python, GCP, SQL, LightGBM, OpenAI APIs, LangChain, Looker Studio, Google Ads, REST APIs.

BU Spark!, Boston, Massachusetts January 2024 - May 2024
Project Manager

- Led three applied ML teams delivering forecasting and optimization tools for civic clients, scoped problem statements, directed model design, and oversaw deployment under noisy, real-world constraints.
- Designed Agile pipelines and stakeholder review loops, accelerating iteration cycles by 40% while maintaining model interpretability and delivery precision.
- Translated high-level, non-technical goals into deployable ML solutions for equity-driven capital planning and governance analytics.
- Tableau, Python, Google BigQuery, Agile, Stakeholder Engagement.

BU School of Public Health, Boston, Massachusetts February 2023 - December 2023
Data Science Research Assistant

- Built time-series forecasting models for IEQ metrics (CO_2, PM2.5, temperature) across 125+ public schools, enabling equity-first prioritization and influencing $MM+ capital allocation.
- Developed an image processing pipeline using OpenCV to extract spatial features from classroom floor plans (e.g., window access, layout density), surfacing disparities linked to environmental quality.
- Designed a real-time air quality dashboard to detect and alert deviations, reducing operational response latency by 35%.
- Python, SQL, R, Python Panel, scikit-learn, OpenCV, Infrastructure Analytics, R Shiny.

Selected Projects

- **Personalized Recommendation Engine:** Built a collaborative filtering system using ALS + MinHashLSH for user-user similarity. Integrated tag-genome regression for cold-starts, validated with Pearson correlation, and evaluated via MAP, NDCG, Precision@100. *Stack: PySpark, Python, GCP (Dataproc).*
- **Transformer Based Summarization:** Fine-tuned a **BART transformer** on news articles with beam search and ROUGE-L optimization. Modeled long-form summarization workflows for content compression and preview generation. *Stack: Python, Hugging Face Transformers.*
- **Semantic Tone Classifier (Humor Detection):** Classified tweet humor using TF-IDF + ensemble models (RF, XGBoost), evaluated with cross-validation and F1. Modeled tone-aware tagging for moderation and content personalization. *Stack: Python, scikit-learn, XGBoost.*
- **Geo-Aware NYC Subway Sentiment Signal Pipeline:** Analyzed 10K+ tweets using RoBERTa + VADER, mapping sentiment to NYC subway ridership by borough. Revealed how public discourse tracks system usage. Stack: Python, RoBERTa, Geopandas, Twitter API.

Skills

- **Programming & Systems:** Python, SQL, C++, R, Rust, Bash, Git, Linux.
- **Machine Learning & Models:** scikit-learn, PyTorch, TensorFlow, LightGBM, XGBoost, Transformers (BERT, RoBERTa, BART).
- **Domain Expertise:** Recommender Systems (ALS, MinHash), NLP, Summarization, Sentiment Analysis, Semantic Classification, Computer Vision, Time Series.
- **Infra & Pipelines:** Spark, Hadoop, Dataproc, Docker, REST APIs, Apache Beam (basic), LangChain, ETL.
- **Cloud & Visualization:** GCP (BigQuery, Looker Studio), AWS, Azure, Tableau, Power BI, Python Panel, Agile Collaboration.