**NEHA SHASTRI**
Boston, MA | +1 (857)-506-3531 | nehags@bu.edu | github.com/nehashastri | linkedin.com/in/nehagshastri/

Analytics graduate student with hands-on experience in Python (5+ yrs), SQL (3+ yrs), and Machine Learning (4+ yrs).

## EDUCATION

**Boston University, Questrom School of Business** Sept 2024 - Jan 2026
M.S. in Business Analytics (Data and Methods) | GPA: 3.89/4.00 Boston, MA
*Coursework: Neural Networks (LLMs, GenAI, multi-modal networks), NLP, Big Data, Statistical Modeling*

**PES University** Aug 2018 - Jul 2022
B.Tech. in Computer Science and Engineering (Machine Intelligence and Data Science) | GPA: 3.75/4.00 Bengaluru, IN
*Coursework: Statistics, Supervised & Unsupervised Machine Learning, Database Management, Causal Inference*

## SKILLS & ACHIEVEMENTS

**Programming & Analysis:** Python (Pandas, NumPy, Statsmodels, HuggingFace, PySpark), SQL, Git, API Development (FastAPI, Flask), Docker, Model Serving

**LLM & GenAI Systems:** Fine-tuning, RAG Architecture, Vector Databases (Pinecone), LangChain, Prompt Engineering, Model Evaluation (ROUGE), Retrieval Strategies

**ML & Deep Learning:** TensorFlow, PyTorch, Transformers (BERT, RoBERTa, Llama), Scikit-learn, Computer Vision (CNNs), Time Series Forecasting, Ensemble Methods (XGBoost, RandomForest)

**MLOps & Production:** MLflow, Weights & Biases, Model Monitoring, Feature Stores, CI/CD, Model Versioning, Experiment Tracking, A/B Testing for Models

**Cloud & Big Data:** Airflow, Prefect, Databricks, GCP (BigQuery, Dataproc, Cloud Functions), AWS (S3, Athena, Bedrock), Apache Spark, Apache Hadoop, YARN

**Visualization & Reporting:** Tableau, Power BI, Streamlit, Looker Studio, Matplotlib, Seaborn, MS Office Suite

**Achievements:** Top 5 among 70 teams in 2025 MinneMUDAC Student Data Challenge
Selected among 100 students to deliver a talk on Artificial Intelligence at the annual Data Symposium.

## WORK EXPERIENCE

Data Science Project Manager @ **BU Spark!** Sep 2025 – Present

- Provide technical mentorship on data engineering, NLP, and visualization best practices across two concurrent projects (20+ students total), guiding the use of Python (Pandas, NumPy, Requests, BeautifulSoup, ArcGIS)
- Guide teams in building ETL and data standardization pipelines with Snowflake, Tableau, and Looker Studio for scalable storage and visualization.
- Facilitate client communications, translating research objectives into actionable technical deliverables—including API integration, troubleshooting blockers.
- Manage GitHub repositories, workflow documentation, and reproducibility standards, overseeing implementation of modular, production-ready outputs aligned with client research goals.

Teaching Assistant @ **Boston University** Sep 2025 – Present

- Handled technical administration of Blackboard, ensuring data integrity and timely updates for 100+ students.
- Conducted statistical analysis in Excel to evaluate student performance trends and support grading decisions.
- Partnered with the professor to refine course materials, providing data-informed feedback on learning outcomes.

Founder & Event Manager @ **Neha Shastri Live** Jul 2022 – Jun 2024

- Worked with 30+ clients, coordinating with marketing and cross-functional teams to deliver large-scale events, increasing audience attendance through targeted regional campaigns.

Web Development Intern @ **Textron Ltd** Dec 2021 – May 2022

- Delivered cross-platform website wireframes (3+ pages) through the SDLC - tracking/completing Jira tickets and pushing code via Azure DevOps repos.

## PROJECT EXPERIENCE

**AI-Assisted Grading Platform for Multi-Modal Submissions** Nov 2025 – Present

*Agentic AI, LLM Pipelines, Reinforcement Learning, API Integration*

- Designed an AI-driven grading engine that applies rubric-based reasoning to essays, diagrams, and spreadsheet submissions, reducing instructor grading load.
- Prototyped multi-modal evaluation flows combining LLMs, structured rubrics, and human-in-the-loop review for transparent and controllable grading outcomes.
- Implemented reinforcement learning from instructor feedback to improve grading accuracy, user trust, and long-term system robustness.
- Integrated the Agentic AI system with Blackboard through API connectors, enabling seamless submission ingestion, scoring, and student feedback delivery.
- Conducted cost–accuracy tradeoff analysis across multiple LLMs to support model selection and operational scaling decisions.
- Worked cross-functionally with educators and EdTech teams to refine product requirements, define pilot success metrics, and support commercialization planning.

**Credit Risk Intelligence Platform for Banks** Sep 2025 – Present

*GenAI, ETL pipeline, Orchestration, Airflow, MLFlow, BigQuery*

- Orchestrated end-to-end ETL pipelines using Airflow and Cloud Functions to automate ingestion of macroeconomic and market data, reducing manual updates by 90% and ensuring daily model-ready datasets.
- Developed BigQuery ingestion and transformation workflows with automated merge logic and validation checks to identify schema drift, inconsistent data types, enhancing data integrity and reliability across daily records.
- Investigated and resolved ETL discrepancies by tracing data lineage and API response logs, designing idempotent DAG logic that skipped redundant runs and cut compute costs by ~30%.
- Engineered feature-ready datasets through interpolation, frequency harmonization, and rolling sentiment aggregations, boosting data completeness by 25% and improving downstream credit-risk model performance.
- Integrated a GenAI system using Gemini to process daily news feeds, generate scenario-based synthetic economic signals, and augment the dataset with forward-looking features for improved risk forecasting.
- Implemented MLflow integration for model tracking, versioning, and performance monitoring with automated retraining triggers based on drift detection
- Documented pipeline structure, variable definitions, and feature transformations, ensuring transparent reproducibility and ease of model maintenance for future data scientists.

**Favorita Stores Sales Forecasting** Apr 2025 – May 2025

*Time Series Forecasting, SARIMA, Prophet, XGBoost, Databricks, Streamlit*

- Automated demand forecasting pipelines for 54 stores and 33 product families, enabling optimized inventory allocation and service delivery decisions through interactive Streamlit dashboards.
- Developed and tested advanced time-series and machine learning models (SARIMA, Prophet, XGBoost), generating predictive insights, achieving 15% MAPE improvement.
- Implemented backtesting and validation workflows with automated daily refresh in Databricks, delivering real-time decision-support tools that enabled data-driven staffing and service delivery optimization.

**Determining Where Eyes Look,** Project Manager Apr 2025 – May 2025

*Deep Learning, Computer Vision - CNN, Fine-tuning, OpenCV, TensorFlow*

- Designed an end-to-end, real-time gaze-detection pipeline using the Columbia Gaze dataset (5,880 headshots, 56 subjects) to perform binary "camera vs. off-camera" classification on live webcam feeds.
- Fine-tuned MobileNet with dropout and L2 regularization, achieving 84% recall while maintaining 10 fps inference on CPU through model quantization.
- Deployed production system using OpenCV with model versioning and fallback logic.
- Audited model decisions with LIME, confirming attention on eye–nose bridge regions; benchmarked alternative detectors and backbones (EfficientNet, ResNet) before selecting the speed/accuracy winner for deployment.

**BBBS (Non-Profit) Dataset: MinneMUDAC Student Data Challenge – TOP 5 TEAM** Mar 2025 – Apr 2025

*NLP, Predictive Modelling (BERTopic, RoBERTa, RandomForest), HuggingFace Transformers*

- Curated 12k+ unstructured interactions—program emails, call transcripts, and survey-free-text; Then built a HuggingFace Transformers preprocessing pipeline (tokenization, embeddings, stop-word & spell-noise cleanup).
- Extracted high-signal features with BERTopic topic modeling and RoBERTa sentiment scores, stacked 3 models into a weighted ensemble that cut baseline RMSE 0.24 → 0.12 (-50 %) on a 5-fold cross-validation hold-out.
- Ranked Top 5 / 70 teams at MinneMUDAC 2024, delivering data-driven recommendations to 25+ executive judges (incl. the CEO), directly influencing the next-cycle mentor-matching service strategy.
- Engineered 10+ NLP features (semantic similarity, topic distributions, sentiment trajectories), improving model interpretability and stakeholder buy-in

**Amazon Reviews: Consumer Sentiment and Trends** Feb 2025 – May 2025

*NLP, Apache Spark, Hadoop, YARN, GCP Dataproc, PySpark*

- Processed 30GB dataset (10M+ product reviews) using PySpark on a multi-node YARN cluster, implementing a distributed NLP pipeline with 8x speedup vs single-node
- Performed topic modeling and sentiment analysis, uncovering 12 key product quality issues, and presented findings to 100+ stakeholders, driving product improvements
- Built predictive model for review ratings using engineered NLP features (topic distributions, sentiment scores, review length), achieving 70% accuracy
- Optimized Spark job configurations and partitioning strategies, reducing processing time from 6 hours to 45 minutes with 40% cost reduction

**Retail-Rocket Customer Behavior Analysis,** Project Manager Feb 2025 – Mar 2025

*Anomaly Detection, K-means Clustering, BERT Transformers, Recommendation Systems*

- Spearheaded anomaly detection to find and remove unusual transactions, resulting in a 50% enhancement in product bundles created based on market basket analysis.
- Performed customer segmentation through batch processing and stabilized K-means leveraging statistical methods, effectively naming two clusters - high-value customers and window shoppers.
- Built hybrid recommendation engine combining collaborative filtering with BERT-based semantic similarity, weighted by recency achieving 0.82 precision@10.

**Experimental Design & A/B Testing for Learning Strategy Optimization,** Project Manager Feb 2025 – Mar 2025

*A/B Testing, CATE Analysis, Statistical Inference, Hypothesis Testing, Python (Statsmodels), Qualtrics*

- Designed a randomized controlled trial with 95 students using blocking to ensure covariate balance, comparing active vs passive learning methods.
- Implemented regression-based cohort analysis and power analysis using t-tests to assess statistical significance and identify heterogeneous treatment effects, revealing +53% participation impact for active learning.

**Microsoft News Dataset (MIND): Predicting News Article Popularity,** Project Manager Nov 2024 – Dec 2024

*Predictive modeling (Model Evaluation, Testing, Validation, Feature engineering, Feature Selection, Hyperparameter Tuning, Ensembling), Prompt Engineering (Open AI API)*

- Implemented parallel processing and feature engineering on user behavior dataset (5M rows to 55K rows), achieving a sixty times reduction in computation time.
- Applied prompt engineering techniques with ChatGPT to summarize articles and extract contextually rich keywords, enabling Google Trends API integration and enhancing features with real-time public interest signals.
- Trained and evaluated a two-step model pipeline combining classification (exceeding null model by five times) and regression models – XGBoost, SVR (lowering RMSE by 10%) to predict article's clicks-to-impressions ratio.
- Presented findings to an audience of 50+ peers, earning recognition for the innovative integration of ML techniques, including ensembling and a custom cost function designed to maximize profit.

**IMDb: Genre, Movie, and Revenue Insights** Nov 2024 – Dec 2024

*SQL, ETL, Bigquery, Tableau, Data Integration & Cleaning, Data Visualization, Dashboard*

- Analyzed 21M+ records across IMDb datasets in BigQuery to uncover audience and revenue trends, providing insights that informed content strategy and customer engagement decisions.
- Designed end-to-end ELT pipelines with advanced SQL (views, window functions, complex joins) and embedded data quality checks, ensuring reliable, scalable frameworks for data-driven service and strategy decisions.
- Deployed 2 interactive Tableau dashboards with 5+ filters (year, region, language), enabling both technical and non-technical stakeholders to explore trends and drive strategic decision-making through digital visualization.