Priyanka Gujar

gujar.p@northeastern.edu | +1(781) 956-3367 | https://www.linkedin.com/in/priyankagujarprofile

SUMMARY

Data Scientist specializing in production-grade ML and GenAI systems, building scalable pipelines and end-to-end solutions across research and applied domains. Experienced in distributed data processing, model evaluation, and applying transformer-based models and knowledge graph workflows to generate reproducible insights.

EXPERIENCE

The Institute for Experiential AI, Northeastern University – Boston, MA Sept 2024 – Present
Machine Learning and Bioinformatics Researcher

- Developed LLM-driven knowledge graphs to extract structured insights from diverse datasets and performed model benchmarking across project workflows, supporting reproducible analysis pipelines.
- Applied transformer-based LLMs to predict chemical properties of enzyme sequences, accelerating candidate screening workflows and reducing manual review effort by 40% across experimental pipelines.
- Designed a modular enzyme function prediction framework, improving classification accuracy by 12% through dimensionality reduction, ensemble modeling, and systematic evaluation.
- Extracted graph-based features and protein embeddings from 122K+ PDB structures using distributed data processing on HPC clusters, enabling scalable experimentation, faster iteration cycles, and reproducible model training.
- Assessed custom protein embeddings against state-of-the-art representations, informing model selection decisions and improving downstream experimental success rates.

Tata Consultancy Services (TCS) – Ahmedabad, India Sept 2021 – Dec 2022
Data Scientist, Assistant Systems Engineer

- Identified high-value customer segments using regression analysis and feature attribution, contributing to a 30% revenue increase and 45% operational efficiency gains across core business workflows.
- Produced real-time KPI reporting through Tableau dashboards and optimized SQL queries, supporting faster executive decision-making and improved visibility into operational performance.
- Streamlined data ingestion and transformation by refactoring PL/SQL workflows, lowering manual rework by 20%.
- Translated analytical findings into business and system requirements in collaboration with engineering and product stakeholders.

PROJECTS

Agentic Data Quality Monitoring & Repair System Dec 2025 – Jan 2026

- Architected a multi-agent framework combining statistical profiling with LLM-driven repair logic, detecting 8+ data issue categories with 95%+ accuracy across heterogeneous datasets.
- Implemented confidence-scored remediation strategies that generated production-ready Python transformations for imputation, validation, normalization, and schema enforcement.
- Established type-safe data contracts using Pydantic, achieving full test coverage and standardized quality scoring across datasets

AI Personalized Learning Assistant with RAG – Intelligent Multimodal Study Companion Jan 2025 – Apr 2025

- Constructed an end-to-end retrieval-augmented generation system, improving answer relevance by 30% compared to baseline LLM inference through grounded context retrieval.
- Enabled multimodal document ingestion with efficient semantic retrieval, maintaining sub-100ms query latency.
- Evaluated system performance using ROUGE, BERTScore, and retrieval metrics, ensuring reproducibility and transparency.

Price Alchemy: End-to-End Production ML Pipeline for Price Recommendation Jan 2024 – May 2024

- Designed & deployed an end-to-end ML pipeline using Airflow, Hyperopt, MLflow, ELK, Docker, and FastAPI, with AWS-compatible architecture, data and model versioning, CI/CD-enabled training and evaluation, achieving 93% predictive accuracy.
- Processed free-text item descriptions with n-gram NLP features, improving model accuracy and enabling segmentation of high-value product cohorts.
- Optimized data preprocessing, training, and monitoring workflows, ensuring reproducible experiments, experiment traceability, and reliable real-time visibility in production workflows.

EDUCATION

Northeastern University, Khoury College of Computer Sciences, Boston, MA
Master of Science, Data Science **GPA: 3.92/4.0**
Courses: MLOps, Supervised & Unsupervised Machine Learning, Data Mining, Computer Vision, Database Management

Savitribai Phule Pune University, Pune, India
Bachelor of Engineering, Electronics and Telecommunication Engineering **GPA: 9.12/10**
Courses: Artificial Intelligence, Machine Learning, Digital Image and Video Processing, Object Oriented Programming

SKILLS

- **Programming & Data:** Python, SQL, R, Matlab, C++, Spark, Data Modeling, Feature Engineering
- **Machine Learning & AI:** Scikit-learn, XGBoost, PyTorch, TensorFlow, Dimensionality Reduction, Ensemble Methods
- **Analytics & Visualization:** A/B Testing, Tableau, PowerBI
- **MLOps:** Airflow, MLflow, Docker, ELK Stack, Model Monitoring
- **GenAI:** Transformers, LLM Evaluation, Embeddings, Model Benchmarking, Retrieval-Augmented Generation (RAG)
- **Data Engineering & Systems**: SQL Optimization, Data Pipelines, Schema Design, Data Validation, HPC Environments