Harshit Soni

hs5666@nyu.edu | 646-203-9176 | New York, NY |Portfolio| linkedin.com/in/harshitksoni | github.com/HarshitSoni1903

Data Scientist with cross-functional experience enabling 5.25× revenue growth, operational efficiency, and policy insights through machine learning and predictive modeling, from analysis to production deployment.

EDUCATION

New York University, New York, NY — *Expected May 2026*
Master of Science in Data Science — GPA: 3.96/4.0
SRM Institute of Science and Technology, India — *May 2019*
Bachelor of Technology in Computer Science and Engineering — GPA: 8.80/10

EXPERIENCE

Machine Learning Researcher, Gyori Lab, Northeastern University [GitHub] — *Sep 2025 – Present*

- Leading a research project on ontology alignment using NLP, representational learning, and semantic search, iterating on evaluation benchmarks with **98% precision (ACM-BCB 2026 Accepted).**
- Fine-tuned a large language model on a knowledge graph using contrastive learning for high-recall retrieval via FAISS (RAG).
- Released as a **plug-and-play end-to-end Python package** for ontology alignment within DARPA-funded MapNet framework.

Data Scientist, The Urban Politics Lab, New York University — *Feb 2026 – Present*

- Proved NYC congestion pricing reduced noise complaints by 70% using statistical analysis (RDD), contributing to a grant proposal on urban noise and transit policy.
- **First-of-a-kind** quasi-experimental geospatial evaluation of board-ups on crime across **6K+ properties and 290K+ incidents**, finding delayed effects contributing to ongoing policy analysis.

Business & Data Analyst, Sunshine Marketing Agencies, India — *Jul 2020 – Aug 2024*

- Established the company's first analytics function, enabling structured data collection across **300 client** accounts while building ETL pipelines, forecasting, and reporting systems; scaled to a team of 5.
- Enabled **5.25x revenue growth ($500K to $2.5M)** over 4 years by improving demand forecasting, inventory allocation, and operations management based on SKU velocity and perishability.
- Reduced product returns by **33% ($40K–50K/yr)** by analyzing seasonal and client-level demand patterns across 700 SKUs.
- Automated client communication via Gmail/WhatsApp/Sheets, **reducing inbound calls by 80%** within three months.
- Built a full-stack order management system (MERN, Firebase) processing 50–100 daily orders across 80+ supplier representatives, reducing **processing time by 50%**.

Systems Engineer, Tata Consultancy Services, India — *Jun 2019 – Jul 2020*

- Identified and automated manual testing chokepoints and SharePoint migration workflows via PowerShell scripts, cutting project delivery time by **35%**; adopted across teams and recognized with **Star Team Award**.
- Created Power BI dashboards tracking operational KPIs for client stakeholders, replacing manual reporting workflows.

PROJECTS

Credit Card Transaction Anomaly Detection [GitHub]

- Detected fraud in credit card transactions (0.17% positive rate, F1 = 0.89) using stacked models (Iso Forest, XGB, LightGBM).
- Deployed the stacked pipeline as a production-grade FastAPI endpoint with Docker for real-time fraud scoring.
- Used SHAP to identify top fraud indicators; tuned thresholds to minimize false blocks on legitimate users.

Movie Recommendation and Segmentation System, [GitHub]

- Segmented 307K users' 27M+ ratings using SQL and PySpark MinHash/LSH, identifying users with identical viewing patterns.
- Showed that similar viewing habits don't mean similar tastes, driving the choice of ALS over heuristic recommendations.
- Tuned PySpark ALS to outperform popularity baselines on MAP and NDCG for top-100 recommendations.

SKILLS

Languages & Libraries: Python, R, SQL, PySpark, scikit-learn, PyTorch, TensorFlow, NumPy, Pandas, Matplotlib, Seaborn
Tools & Infrastructure: Git, Docker, AWS, GCP, Tableau, MongoDB, Snowflake, dbt, Excel, GitHub Actions
Methods: Big Data, Machine Learning, Statistical Modeling, Exploratory Data Analysis, Feature Engineering, A/B Testing, Forecasting, Causal Inference, Deep Learning, Data Visualization, Model Validation, Optimization, Time-Series Forecasting