RISHABH PATIL

+1 929-424-7773 • rbp5812@nyu.edu • linkedin.com/in/rishabhbhaskarpatil • github.com/rishswish • rishswish.github.io

EDUCATION

New York University, Center for Data Science September 2024 – May 2026

M.S. in Data Science GPA 3.86/4.0

Coursework: Natural Language Understanding, Big Data, Machine Learning, ML in Finance, Data Engineering

University of Mumbai July 2020 – May 2024

B.Tech in Computer Science and Engineering (Data Science) with Honors in Computational Finance GPA 3.96/4.0

Coursework: Time-Series,Cloud Computing,Deep Learning,Database Management,Computer Vision,Reinforcement Learning

SKILLS

Technologies: Python, R, SQL, Git, TensorFlow, PyTorch, SciKit-Learn, OpenCV, PySpark, Hadoop, Dask, Kafka, Azure Synapse, CUDA, Cassandra, Kubernetes, Snowflake, Jenkins, Docker, PowerBI, Tableau, AWS, MongoDB, MS Excel, RedShift

PROFESSIONAL EXPERIENCE

Muck Rack, New York, USA | ML Engineer Intern August 2025 - Current

- Developed and Operationalized an HTML quality detection system as a versioned internal Python package, enabling reuse across Muck Rack teams.
- Implemented **CI-ready unit tests** and a **regression-test harness**, ensuring production performance stayed within **5 pp** of local benchmarks.
- Rolled out a daily monitoring pipeline (**S3 → inference → Snowflake**) and optimized inference to process **10,000+ HTML pages.**

Solar Secure Solution, India | Data Analyst Intern February 2023 - April 2023

- Orchestrated a real-time IoT telemetry analytics pipeline using **Pandas, Plotly, and Streamlit** to visualize latency, packet loss, and device uptime across thousands of nodes—cutting incident response time by **40%** and enabling preventive monitoring.
- Designed predictive insights via time-series modeling (**ARIMA, rolling averages**) to flag high-risk devices with **78% precision**, helping network teams preempt failures and optimize infrastructure planning.

Acmegrade Pvt Ltd, India | Machine Learning Intern July 2022 - September 2022

- Engineered a Python-based **Computer Vision + NLP pipeline** using **Azure OCR** to auto-extract key data fields from PDFs, screenshots, and log files—**reducing processing time by 25%** while increasing downstream analytics throughput.
- Devised **prompt-engineering modules** that **auto-generate** charts, slide decks, and multilingual summaries that **eliminated 75%** of manual review effort and **expanded** cross-team visibility.

ACADEMIC PROJECTS

Progressive Learning in LLMs with Structured Grammar Books| Github January 2025 – May 2025

- Compiled a **345-lesson** curriculum with **Tesseract OCR** + **Stanza**, generating **1.7K** syntax feature vectors (POS/DEP/NER).
- Implemented syntax-augmented Transformers (**SyntaxGPT, SyntaxT5**) in **PyTorch**/**Hugging Face** using token+syntax embeddings; cut training **2.5 days→3 hours** (**-95% compute**) and achieved **87%** on **TREC** with **52% faster inference** (**236s→114s**).

Personalized Recipe Recommendation System| Github July 2023 – May 2024

- Developed a **Flask** app powered by a **GPT-4 + text-embedding-ada-002 + LanceDB** RAG pipeline, improving **recipe-match accuracy by 40%** and increasing engagement by **35%** across **10K+ sessions**.
- Integrated allergy-aware cosine scoring + real-time feedback to reduce irrelevant suggestions by **25%** and raise user satisfaction by **50%**, mitigating cold-start and hallucinations.

Driver Drowsiness Detection System| Github January 2022 – January 2024

- Trained dual **YOLOv5** detectors (eye-closure & yawning) on **1.2K+ annotated images**; achieved **85% accuracy** and improved real-time alert response by **30%** using probability scoring + voice alarms.
- Incorporated CNN-based fusion to improve robustness across **1M+ video frames** (low-light/occlusion), reducing false positives by **20%** and boosting accuracy by **15%**; awarded **3rd Prize** (ICDMAI 2024) and published in Springer LNNS 998. *Article*.

MovieLens Recommendation & Segmentation| Github January 2025 – May 2025

- Established a terabyte-scale **PySpark + Hadoop** pipeline on the full **MovieLens corpus (330K users / 86K movies)**, enabling distributed analytics and user segmentation via **MinHash + LSH** with sub-second latency.
- Delivered two recommenders: a **Spark ALS** model with **+30% Precision@100** in high-coverage cohorts, and a **bias-corrected popularity model** for 90% sparse data, yielding **+66% MAP** and **20× ALS performance**.