Nikhil Soni

Brooklyn, NY | nikhilsoni700@gmail.com | +1 9342331695 | linkedin.com/in/nikhilsoni15/ | nikhilsoni15.github.io

Education

New York University, New York, NY — May 2025
Master of Science in Computer Science (*Recipient of Merit-based scholarship)* — GPA: 3.97/4.00

- **Relevant coursework:** Data Structures and Algorithms, Data Science for Business (NYU STERN), Machine Learning, Artificial Intelligence, Deep Learning, Big Data Analytics, Database Systems, Web Search Engines, Programming Languages

Manipal University, Jaipur, India — Aug 2019 - Jul 2023
Bachelor of Technology (BTech) in Computer Science — CGPA: 9.16/10.00

Technical Skills

Languages: Python, SQL, C, C++, Java, HTML, CSS, JavaScript

AI / ML: Scikit-learn, NumPy, Pandas, TensorFlow, PyTorch, Hugging Face, LLM fine-tuning, RAG, NLP, NLTK, Streamlit

Cloud / DevOps: AWS (EC2, S3, Lambda, API Gateway), GCP, Azure, Docker, Kubernetes, Kafka, Airflow, Spark Streaming

Databases & APIs: Advanced SQL (Window Functions, Indexing, Query Optimization), FastAPI, Qdrant, MongoDB, Redis

Experience

Graduate Teaching Assistant - Deep Learning, *New York University* – New York, NY — Jan 2025 – May 2025

- Orchestrated and hosted 2 Kaggle-style competitions to evaluate student solutions in NLP, Generative AI and Transformers
- Mentored 400+ students through office hours, clarifying concepts in diffusion models, RL, and advanced deep learning

AI/ML Intern, *Emerson* – Pune, IN — Jun 2024 – Aug 2024

- Architected an end-to-end LLM-based tool to automate validation of DeltaV system control reports, reducing manual effort
- Elevated recognition accuracy to 91% by applying chain-of-thought prompting on T5 and BERT, increasing tool efficiency
- Streamlined data pipeline, processing 10,000+ text files saving around 25–30 human hours weekly when performed
- Collaborated to fine-tune LLMs, reducing model training time by 25% and improving alignment with domain-specific data
- Automated log and report generation with highlighted error sections, enhancing stakeholder data traceability

Data Science Intern, *Junglee Games* – Gurugram, IN — Jan 2023 – Jul 2023

- Extracted and analyzed Fraud users data using SQL and Python-based EDA to build and optimize predictive models at scale
- Led testing and deployment of the "Problem Gamer" model to catch game addicts in a pool of 100 million users
- Optimized deployment and monitoring through MLOps pipelines using AWS Lambda, reducing model update time by 18%
- Replicated a CNN research paper to calculate players Rummy skill score to predict game drop decision with 82% precision

Software Developer Intern, *Hewlett Packard Enterprise* – Chandigarh, IN — Jun 2022 – Jul 2022

- Developed a full-stack application with Django backend and HTML, CSS, JavaScript frontend for intra-team issue reporting
- Provisioned the system on AWS using EC2 instances and VPC, ensuring scalability, security, and high availability

Projects

CrisisCast: Real-Time Crisis Detection & Monitoring | *PySpark, LLM, Kafka, Qdrant, MongoDB* [Github]

- Built end-to-end real-time emergency detection system by ingesting Reddit data using Kafka and Spark Structured Streaming
- Integrated a locally hosted LLM-based classifier to tag posts by crisis type, storing enriched metadata into MongoDB
- Embedded 1,000+ Reddit posts into vector space using Sentence Transformers and stored semantic representations in Qdrant
- Designed an interactive Streamlit dashboard for real-time crisis trend monitoring and semantic search across incoming posts

GenVision: Personalized Image Generator | *Diffusion XL, DreamBooth, LoRA, Flask, Gradio* [App|Code]

- Constructed full-stack app fine-tuning Stable Diffusion XL with DreamBooth and LoRA for personalized prompt-based images
- Created a Flask backend and Gradio-based frontend for real-time, user-specific image generation along with a feedback slider
- Applied optimizations like gradient checkpointing and mixed-precision training, to enhance performance on limited GPUs

Rent Raja - NYC Rental Price Prediction | *Scikit-learn, Flask, Dash, LLM, APIs* [Demo|Code]

- Crafted predictive ML model to estimate rental prices by mining and processing 300,000+ property listings from various APIs
- Devised hybrid classification-regression pipeline, 81% classification accuracy on 3 bins, reducing RMSE from $3,000 to $300
- Engineered 10+ predictive features along with Dash + Flask dashboard and LLM-generated reports for broker pricing insights

Web Search Engine | *TensorFlow, Python, C++* [Github]

- Implemented a web crawler and inverted index system, processing 12,000+ web pages to enable large-scale data retrieval
- Refined storage with VarByte compression and index sharding, reducing overhead by 30% and improving query speed
- Established a query processor using BM25 scoring, designing ranking algorithms to handle complex queries with precision
- Sharded the index using K-Means, MiniBatch K-Means, and GMM clustering on 3M-document vectors

OopsFix: 311 Service Optimization for NYC Boroughs | *Pandas, Numpy, Scikit-learn, ETL* [Github]

- Consolidated scalable ETL pipelines to process 34 million records, integrating multimodal data for efficient processing
- Fine-tuned ensemble learning models, achieving 87% accuracy in predicting borough-specific service delays
- Conducted spatiotemporal analysis to identify inefficiencies, integrating multimodal data to optimize resource allocation