Harshit Joshi

(+91) 8826252219 | ✉ harshith.j.2003@gmail.com | New Delhi

Harshit | harshit1912003 | HarshitJos2003 | harshit1912003.github.io

Overview

Mathematics graduate with proven experience in developing end to end AI/ML models and implementing optimization solutions across diverse projects. Eager to leverage this strong analytical foundation for pioneering research in the field of stochastic modeling and optimization.

Experience

- **Research Internship** *Dec, 2024 - Jan, 2025*
  Dr. Aparna Mehra, HOD, Mathematics IIT Delhi New Delhi
  - Explored various Non-Homogeneous RDM models using distinct output subgroups; implemented Cook et al.'s framework in Python to extend DEA for general non-homogeneous datasets.
  - Replicated results from Pooja Bansal's paper "Non-homogeneous DEA approach in the presence of negative data".
- **Machine Learning Intern** *May, 2024 - July, 2024*
  Swachh.io New Delhi
  - Built a YOLOv8-based vehicle emission detection model with ANPR for Indian vehicles; trained on specialized web sourced data. Achieved 89.5% precision, 80.7% recall, 88.4% mAP50, and 62.7% mAP50-95.
  - Scraped classroom counts from appx 8,500 schools across Karnataka using Python and BeautifulSoup.
  - Designed and implemented statistical methods to assess product filtration efficiency with subsequent data analysis.

Education

- **Indian Institute of Technology Delhi** *July, 2023 - May, 2025*
  M.Sc. in Mathematics New Delhi
  - CGPA: 7.8/10
- **Hans Raj College, University of Delhi** *Aug, 2020 - June, 2023*
  B.Sc. in Mathematics New Delhi
  - CGPA: 8.8/10
- **Khaitan Public School** *CBSE*
  Schooling
  - Senior School Certificate Examination (12th): 95.33%
  - Secondary School Examination (10th): 96.20%

Research Outputs

S=In Submission, T=Thesis

[S] **Novel DEA Models for Non-Homogeneous DMUs with Negative Data** (Sumbitted - EJOR)
3rd Author

- Developed novel Data Envelopment Analysis (DEA) models to assess the relative efficiency of non homogeneous Decision Making Units (DMUs) operating with negative input-output data.
- Introduced range directional measure (RDM)-based convex DEA and non-convex Free Disposal Hull (FDH) models, specifically formulated to address negative input-output data within non-homogeneous DMU contexts. To mitigate overfitting prevalent in traditional models, a novel RDM-based Efficiency Analysis Tree (EAT) model was developed for non-convex technologies, alongside an RDM-based Convexified EAT (CEAT) model for convex technologies.
- Demonstrated, via rigorous comparative analysis and statistical hypothesis testing, that the proposed RDM-based EAT and CEAT models exhibit superior robustness in efficiency assessment by effectively circumventing the overfitting of the efficient frontier.
- Implemented a benchmarking methodology to identify optimal peer sets for inefficient DMUs at a granular node level, deriving actionable recommendations for performance optimization.

[T] **Non-Convex Data Envelopment Analysis** (Awarded JRC Award for best thesis (MSc))
Supervisor: Dr. Aparna Mehra, Co-Author: Ananya Sharma
This thesis presents a comprehensive study of fundamental and advanced models in Data Envelopment Analysis (DEA). Key contributions include:

- Detailed study and comparative analysis of classical DEA models (CCR, BCC, Additive, SBM, Modified SBM) and Free Disposal Hull (FDH) models (CRS/VRS, Input-/Output-oriented), including economic interpretations, applications, and statistical validation.

- In-depth exploration of Efficiency Analysis Trees (EAT), a novel method adapting Classification and Regression Trees (CART) for production frontier analysis under Free Disposal assumptions, addressing overfitting, supported by extensive Monte Carlo simulations.
- Development of an open-source Python library for efficient computation of DEA and FDH models using Gurobi, including support for non-homogeneous data analysis.
- Design of a user-friendly Graphical User Interface (GUI) to simplify data input, model execution, and result visualization, aimed at encouraging wider research adoption.

All source code and the GUI are available on GitHub: `link`.

PROJECTS

- **Chessboard FEN Estimation using novel features**
  Deep CNNs, Feature Engineering, Regression Analysis []
  - Designed a Deep CNN architecture to approximate Stockfish Evaluations from a given FEN state of a chessboard, achieving a final validation Mean Absolute Error of 152 centipawns.
  - Developed novel features derived from linear regression parameters of board state characteristics, which significantly improved the model's validation performance by about 100 centipawns.
  - Deployed the trained model through a user interface, accessible at mtl782-chessapp.streamlit.app,
- **Text Guided Image Clustering**
  Deep ImageNet Models, Transfer Learning, Sentence Embeddings, Captioning, Visual Question Answering []
  - Performed comprehensive image clustering on the Food-101 dataset using a multi-modal approach, evaluating classical features (SIFT, Canny Edge, Color Histograms, LBP, HOG) alongside deep features derived from pre-trained CNNs (ResNet-50, EfficientNet-B0).
  - Explored text-guided feature representations by employing Vision-Language Models: BLIP for caption generation and ViLT for Visual Question Answering subsequently using SBERT to get associated feature embeddings.
  - Fine-tuned the BLIP model and utilized SBERT to generate semantically rich text embeddings, which, when combined with DBSCAN, yielded an ARI of 0.8041.
- **Clustering using ToMATo Algorithm**
  Topological Data Analysis, Density-based Clustering, Persistent Homology []
  - Explored Topological Mode Analysis Tool (ToMATo) for clustering high-dimensional point clouds using topological persistence to overcome and merge noise artifacts detected by traditional density based clustering methods.
  - Performed extensive testing of ToMATo algorithm across various data dimensionalities achieving ARI of 0.78 for MNIST digits dataset when compared to k-Means (0.52)
- **Multi-Category Text Classification**
  C-(Bi)LSTM, (CBOW, Word2Vec, Dynamic Meta) Word Embeddings, Self-Attention, Positional Encoding, Transformers []
  - Designed and evaluated multiple deep learning models for multi label text classification, including C-LSTM and C-BiLSTM architectures with custom-trained CBOW word embeddings; incorporated self-attention mechanisms to enhance feature representation.
  - Developed dynamic meta word embeddings by combining custom CBOW with pre-trained Word2Vec vectors; experimented with model variants integrating meta-embeddings and self-attention as well.
  - Implemented Transformer Encoder-based models with variations in positional embedding strategies and encoder depth; conducted comparative analysis across architectures based on test accuracy.
- **N-FLP Robust Linear Regression**
  Robust Estimators, Econometrics, Monte Carlo Testing []
  - Implemented the Normal-Filtered Log-Pareto (N-FLP) mixture model for robust linear regression, a methodology proposed by Alain Desgagné (2019) to achieve efficient and robust estimation in the presence of heavy-tailed error distributions.
  - Applied the N-FLP model to the IPUMS USA 2019 dataset to fit a Mincer wage function, leveraging the model's inherent outlier detection mechanism to identify and analyze economic outliers, providing insights into deviations from typical wage determinants.
- **S&P 500 Diversification**
  Portfolio Optimization, Hierarchical Clustering, Quadratic Programming, Risk Management []
  - Designed a novel portfolio optimization framework for the S&P 500 using hierarchical clustering to model asset dependencies and mitigate concentration risk arising from sectoral or individual exposures; formulated the allocation as a Quadratically Constrained Quadratic Program (QCQP) minimizing the maximum intra-cluster variance across hierarchical levels.
  - Introduced an a priori constraint reduction method leveraging inter-cluster covariance structure, reducing constraints from 501 to 9, enabling tractable optimization; the resulting portfolio outperformed Hierarchical Risk Parity (HRP) with a Sharpe Ratio of 1.15 vs 0.91 over a 10-year backtest.

- **Axelrod’s Tournament**
  Game Theory, 2-Player-0-Sum Games []
  - Built a Python simulation of the Prisoner’s Dilemma, featuring multiple strategies such as Tit-for-Tat, Random, Friedman, Joss, and Harrington with flexibility for users to customize and add their own strategies.
  - Implemented a tournament framework to evaluate the effectiveness of strategies, using cumulative payoff tracking and graphical result representation

Skills

- **Languages & Databases:** Python, R, C/C++, SQL, NoSQL
- **ML & LLMs:** PyTorch, TensorFlow, Hugging Face, LangChain, LlamaIndex, RAG, scikit, Spark, Hadoop
- **Deployment & Tools:** AWS, GCP, Azure, Docker, FastAPI, Git, Vector DBs (FAISS, Pinecone)

Extra Curricular Activities

- **Co-Author:** In Search of the Perfect Story (Quill Club Writers, 2018)
- **Volunteer:** Mathematics Society, IITD (2023–2024)
- **Certifications:** Operations Research (1, 2 & 3) – NTU; ML in Production; OpenCV Bootcamp

Relevant Coursework

MTL505 (Computer Programming), MTL502 (Linear Algebra), MTL508, 32357616 (Mathematical Programming), ELL784 (Introduction to Machine Learning), MTL601 (Probability and Statistics), MTL732, 32357614 (Financial Mathematics), HSL613 (Econometrics), AIL721 (Deep Learning), MTL782 (Data Mining)