Era Parihar

+1 (734) 412-9886 | eraparihar1@gmail.com | linkedin.com/in/eraparihar/ | https://erapariharr.github.io/

EDUCATION

University of Michigan - School of Computer Science | Ann Arbor, MI ***August 2023 - May 2025***

Master of Science in Data Science

Relevant Coursework - Statistical Learning I: Regression, Natural Language Processing, Large Language Models, Machine learning, Data Manipulation

Birla Institute of Technology and Science ***September 2016 - August 2020***

Bachelor of Engineering in Computer Science

Relevant Coursework - OOP (Java), Data Structures & Algorithms, Database Systems, Mathematics, Neural Networks, Computer Architecture, GIS

WORK EXPERIENCE

Data and AI Associate | *United Nations* ***September 2024 - December 2024***

- Fine-tuned ClimateBERT & Logistic Regression on humanitarian text data to classify disaster type & region, improving accuracy to 85%.
- Designed and deployed an end-to-end RAG pipeline on Azure, combining BERT embeddings with vector search to generate interactive, cited country reports.
- Engineered metadata features (e.g., region, climate zone) and built regression models to score climate risk levels across regions, supporting prioritization of aid deployment; evaluated performance using Precision, Recall, and F1, and worked with cross-functional policy and engineering teams to ensure outputs aligned with field priorities.

Data Scientist - Growth & Marketing Analytics | *Deriv Limited* ***April 2021 - May 2022***

- Built and deployed Random Forest and Linear Regression models to predict affiliate performance KPIs (e.g., revenue contribution, retention likelihood), improving targeting strategies and increasing overall retention by 18% in 3 months.
- Did feature engineering on user behavior and marketing data; evaluated models using R^2, RMSE, AUC, iterated based on CV results.
- Developed a Top-K recommendation system for Affiliate Managers, boosting user engagement by 33% via targeted marketing strategies.
- Ran A/B tests (with power analysis and significance checks) to evaluate UI and email changes, leading to higher click-through rates.
- Built ETL pipeline using Apache Spark, Airflow, and PostgreSQL, reducing processing time by 25% and enabling real-time CLV tracking.

Junior Data Scientist – ESG & Forecasting | *SCS Enviro Services Pvt. Ltd* ***July 2020 - March 2021***

- Built ARIMA and linear regression models to forecast air quality ($PM2.5$, NO_2), driving 15% more accurate forecasts used in ESG dashboards to guide environmental strategy and regulatory compliance.
- Created ETL workflows using Python and SQL to prepare environmental data and power forecasting dashboards.
- Designed end-to-end ML pipelines in Python and SQL for data ingestion, preprocessing, modeling, and reporting; automated outputs via Power BI/Tableau dashboards.

Computer Vision Intern – Autonomous Systems | *Sentient Labs* ***April 2020 - July 2020***

- Built an obstacle detection system using object detection (YOLO) and semantic segmentation techniques (e.g., bounding box detection and pixel classification), tailored for aquatic robot navigation in dynamic waterway conditions.

RESEARCH EXPERIENCE

Summer Researcher | Carnegie Mellon University - MSLP Group | Pittsburgh, PA ***May 2024 - July 2024***

- Deep-dived into LLM pre-training across text (MLM, CLM, PLM, SOP) and speech (MAM, CPC) tasks to identify initial learning stages
- Conducted speech & multimodal experiments using contrastive learning and acoustic modeling to enhance performance under Prof. Raj.

Summer Researcher | University of Michigan - The Brunaugh Lab | Ann Arbor, MI ***July 2024 - August 2024***

- *A Computer Vision Framework for Dissolution Profiling of Microparticles* - Engineered a computer vision pipeline using image preprocessing, morphological operations, and contour detection to quantify dissolution kinetics from time-lapse microscopy; extracted time-series features to model drug-specific degradation trends.

PROJECT EXPERIENCE

MedQuery: Evidence-Based Clinical Decision Support

- Developed a RAG-based Clinical Decision Support System combining chain-of-thought reasoning with verifiable citations from PubMed. Integrated real-time query resolution and prompt engineering to enable explainable medical recommendations.

Answer-Aware Question Generation

- Fine-tuned T5-large and BART-large on a multi-dataset corpus (SQuAD, AdversarialQA, MS MARCO) with paraphrased question variants to improve linguistic diversity and robustness. Designed evaluation pipeline using BLEU, ROUGE, METEOR, BERTScore, and GPT-3.5 to assess grammatical correctness, semantic alignment, and model performance

Register-Augmented LLM Fine-tuning

- Co-Developed a “register-augmentation” technique for transformer models (e.g., BERT), inserting specialized tokens during fine-tuning to improve QA performance; used interpretability methods (Integrated Gradients, Layer-wise Relevance Propagation (LRP)) to show enhanced focus on task-relevant context, boosting F1 and ExactMatch scores.

Economic Influences on the News Dynamics Using Statistical Modeling

- Analyzed 2012–2018 news text using token frequency, LDA topic modeling, and time-series regression to study economic effects on news cycles. Applied Spearman correlation and OLS regression, revealing limited correlation relationship between economic indicators and media topic shifts.

Ann Arbor Water Production Forecasting

Using Voting Regressor for Time-Series Prediction

Pothole Prediction for Chicago

- Applied LightGBM algorithm to predict the number of monthly potholes in various areas using the dataset from the Chicago Data Portal between 2011-2016, aiding in efficient city planning. Secured first rank in the Kaggle competition among 80 contestants

SKILLS

Machine Learning & Modeling: Linear & Logistic Regression, Random Forests, XGBoost, Time Series Forecasting (ARIMA, SARIMA), Clustering (K-Means, DBSCAN), Topic Modeling (LDA), Gradient Boosting, Recommendation Systems, A/B Testing, Hypothesis Testing

Natural Language Processing & LLMs: BERT, T5, BART, LangChain, Transformers, Text Classification, Tokenization

Programming & Libraries: Python, SQL, R, Java, Pandas, NumPy, Scikit-learn, PyTorch, PySpark, NLTK, SpaCy, Streamlit, PostgreSQL

Data Engg and Cloud Platform: Apache Spark, Hadoop, Apache Airflow, Docker, Kubernetes, MySQL, Snowflake, BigQuery, AWS, Azure

Visualization & BI Tools: Tableau, Power BI, Matplotlib, Seaborn, Google Looker

Statistics: Descriptive & Inferential Statistics, Hypothesis Testing, Confidence Intervals, Regression Diagnostics, ANOVA, A/B Testing