

ROHIT MACHERLA

📞 940-594-9613 ✉️ rohitmacherla125@gmail.com 🌐 [in/rohitmacherla125](https://in.rohitmacherla125) 🐙 [RohitMacherla3](https://github.com/RohitMacherla3) 📁 [Portfolio](#)

Work Experience

Rutgers University, Graduate Research Assistant | [Git](#)

Oct 2022 – Dec 2023

- Analyzed media data using **topic modeling** to uncover hidden narratives. Achieved **10x clustering speedup** with FAISS in KMeans, delving into DBSCAN, DP-Means, and ultimately opting for BERTopic, uncovering 150+ clusters
- Optimized data integrity through data standardization across 3 sources and preprocessing using NLP techniques resulting in a **reduction** of the data **by 30%** and **removal** of URLs, HTML tags, and emojis **by 99%**
- Analyzed health datasets, designed and implemented Python algorithms to calculate gene-drug interactions, and **identified the top 10%** cases of interest based on estimated **statistical** parameters
- Enforced parallel processing over 64 cores of a remote server, resulting in an **80% reduction** in execution time

Omdena, Machine Learning Engineer | [Git](#)

May 2023 – Aug 2023

- Collected and curated crowdsourced data from over 75+ contributors, conducted EDA, and employed data cleaning and statistical imputation techniques to **enhance data quality** by achieving a **98% completion rate**
- Engineered a **recommendation system** that leveraged content-based, collaborative filtering and NLP techniques. Explored matrix factorization and neural networks, to achieve a **94% f1-score**
- Implemented an ensemble model, to enhance the **click-through rates by 33%**
- Deployed the models to AWS utilizing Streamlit and FastAPI for users to interact and test as a POC

Deloitte Consulting, Data Engineer

Aug 2020 – Jun 2022

- Led** a team of 4, seamlessly integrated **Databricks** with Collibra Catalog using JDBC simba spark driver. Automated metadata ingestion for 260+ schemas using Python scripting and Tidal jobs to **reduce manual effort by 99%**
- Developed **SQL queries** to extract data from MySQL, Oracle, and PostgreSQL databases, optimizing efficiency, and seamlessly exposed the results as APIs using MuleSoft proxy for enhanced accessibility and integration
- Engineered an **ETL** pipeline for data processing automation of Qlik Sense data into Informatica Cloud (IICS) and Collibra via REST API calls using Unix Script, parallelized the ingestion for a **66% time reduction**
- Designed **Power BI** Dashboards through Collibra APIs, showcasing asset metrics and metadata completeness, driving a **30% accuracy awareness boost**. Incorporated **3-layer drill-through** for enriched asset lineage comprehension
- Facilitated and actively contributed to the successful execution of **5 production releases**, demonstrating expertise in deploying and maintaining data engineering solutions to ensure operational stability

Technical Skills

- Domain Expertise:** Data Analytics, Data Engineering, Data Mining, Machine Learning, Deep Learning, Statistics, A/B Testing, MLOps, Natural Language Processing, Computer Vision
- Proficient:** Python, SQL, Pandas, Numpy, Sklearn, TensorFlow, PyTorch, NLTK, OpenCV, Statistics, Matplotlib, GCP([Certified Data Engineer](#)), BigQuery, Databricks, Collibra, ETL, Informatica Cloud
- Worked with:** R, Unix, MongoDB, Power BI, PySpark, SageMaker, Docker, FastAPI, Streamlit, LLMs,
- Basics:** Airflow, Kafka, MLflow, Tableau, spaCy, AWS

Education

Rutgers University-New Brunswick, NJ

Master of Science in Data Science

Graduating: May 2024

GPA: 3.65/4

National Institute of Technology Kurukshetra, India

Bachelor of Technology

Aug 2016 – May 2020

GPA: 8.47/10

Projects

Human Emotion Detection | [Git](#) | [Demo](#)

Apr 2024

- Implemented image classification using ResNet50 **CNN** architecture building it from scratch to achieve an f1-score of 78.6%
- Performed **Transfer Learning** on EfficientNetB4 to **improve** the score to **83.3%**. Fine-tuned the **Vision Transformer** model to obtain an accuracy and f1-score of **90%**
- Quantized** the model to **reduce the size by 92%** and deployed the model as a web app using **Streamlit**

Customer Churn Prediction | [Git](#) | [Demo](#)

Mar 2024

- Performed survival analysis to identify the likelihood of churn over time and built a classification model, selecting the best-performing model from various **tree-based** models to predict customer churn
- Implemented **SMOTEEN** and **hyperparameter** tuning to optimize the AUC score by 15% to **98.9%**, and deployed the **XGBoost** model as a user-friendly web application using Streamlit

Text Summarization - NLP | [Git](#) | [Demo](#)

Dec 2023

- Performed text summarization on wikiHow dataset from Hugging Face using pre-trained **BART** and **T5 LLMs**
- Prompt Engineering** techniques were used on the input text to **improve** the BLEU score **by 10%**
- Adopted simple and **LoRA fine-tuning** techniques to improve the **performance** on BLEU score **by 209%**