ROHIT MACHERLA

940-594-9613 rohitmacherla125@gmail.com in/rohitmacherla125 RohitMacherla3

Education

Rutgers University, New Brunswick, NJ **Sep 2022 – May 2024**
Master of Science in Data Science *GPA: 3.69/4*

Coursework: Regression & Time Series Analysis, Data Structures & Algorithms, Probability & Statistics, Data Mining, Statistical Modeling & Computing, Database Management, Statistical Software (NLP), Statistical Learning (DL)

National Institute of Technology Kurukshetra, India **Aug 2016 – May 2020**
Bachelors of Technology in Electrical Engineering *GPA: 8.47/10*

Technical Skills

Tools and Languages

- **Proficient**: Python, MYSQL, Pandas, Numpy, PyTorch, Scikit-learn, NLP, NLTK, Matplotlib, Seaborn
- **Worked with**: Unix, R, MongoDB, Tableau, Surprise, Tensorflow, spaCy, FastAPI, Streamlit, LLMs, Transformers

Cloud Technologies: GCP(Certified Data Engineer), Databricks, BigQuery, Collibra, Informatica Cloud (IICS)
Data Science Skills: Machine Learning, Deep Learning, Generative AI, Recommendations Engines, A/B Testing

Work Experience

Graduate Research Assistant **Oct 2022 – Dec 2023**
Rutgers University | Topic Modeling, NLP, Statistical Analysis | Git *New Brunswick, NJ*

- 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%**
- Achieved **10x clustering speedup** with FAISS in KMeans, delving into DBSCAN, DP-Means, and ultimately opting for BERTopic, uncovering 150+ clusters
- 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

Machine Learning Engineer **May 2023 – Aug 2023**
Omdena | Recommendation Systems, Predictive Analysis | Git *Remote*

- Collected and curated crowdsourced data from over 75+ contributors, conducted EDA, and employed advanced data cleaning and 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

Data Engineer **Aug 2020 – Jun 2022**
Deloitte Consulting *Hyderabad, India*

- **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%**
- Collaboratively linked Qlik Sense data with Collibra Catalog via REST API calls using Unix Script and Curl utility, involving 3 team members. Handled about 1000 Applications with 15000 Sheets
- Efficiently processed, transformed, and loaded Qlik Sense data into **Informatica Cloud** (IICS) and Collibra, parallelizing 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

Projects

Text Summarization | *NLP, LLMs* | *Git* **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%**

Twitter Search Application | *MongoDB, MYSQL, Caching* | *Git* **Apr 2023**

- Extracted and transformed user and tweet data from JSON files, succeeded with **100% accuracy** in storage into MYSQL and MongoDB, leveraging analytics tools for EDA and sentiment analysis
- Developed search functionalities with custom **caching** techniques and attained a **500-fold enhancement**

Air Quality Prediction | MICE, Bootstrapping, Elastic Net Regression| *Git* **Apr 2023**

- Discerned missing data patterns using the **MCAR** test and created **5 imputed** datasets with MICE, selected the best-performing dataset using the **PMM** technique in R Studio
- Executed **bootstrapping** with 1000 samples to estimate bias and variability, and constructed a robust model using the **elastic net** regularization to **decrease the error by 10%**