RIJUL SHERATHIA

rsherathia@ucsd.edu | San Diego, California | (619) 953-7403 | LinkedIn | GitHub

EDUCATION

MS, Data Science | University of California, San Diego | 3.93 GPA **June 2024**

- *AWS Certified Cloud Practitioner*
- *Teaching Assistant for Graduate Courses:* ***Statistical Models*** *and* ***Fraud Analytics***

B.Tech, Computer Science and Engineering | MIT World Peace University, Pune | 3.9 GPA **June 2021**

- *Data Science A-Z (Udemy), Machine Learning A-Z (Udemy)*

SKILLS

Languages	:	SQL, Python (Pandas, NumPy, Sklearn, Matplotlib, Tensorflow, Keras, PyTorch), R, Matlab, Java, C++
Databases	:	PostgreSQL, Redshift, Oracle, MySQL, NoSQL, MongoDB, Neo4j, VectorDB, CosmosDB, SQL Server
Tools/Skills	:	Apache (Kafka, Airflow, Spark, Hadoop, Beam), Tableau, Power BI, Alteryx, DBT, Dataflow, Time Series Analysis, PCA, EDA, Hypothesis Testing, Regression Analysis, SVN, Github, SnapLogic, Erwin, ML Algorithms, CNNs, LLMs, GANs, VAEs
Cloud	:	AWS (RDS, Redshift, DynamoDB, S3, Lambda, EC2, Glue, Athena, SNS, CloudFormation, IAM, SageMaker)

WORK EXPERIENCE

Data Science Researcher | iNetMed Lab | San Diego, CA (June 2023 - Dec 2023)

- Utilized **R** for **Exploratory Data Analysis (EDA)** in genomics, applying statistical techniques like **PCA, t-test, ANOVA, regression analysis** to understand variables, unveiling patterns and enhancing gene network prediction by 15%.
- Engineered automated pipeline to process **RNA-Seq data** in **R** using (DEseq2, Bioconductor) for **normalization, differential expression analysis**, generating customized plots, and facilitating informed decision-making.

Machine Learning Researcher | NCMIR Lab | San Diego, CA (June 2023 - Dec 2023)

- Applying image segmentation **(U-Net Attention, Mask R-CNN)** on EM brain cells using **TensorFlow**, achieving 76% model accuracy.
- Implemented preprocessing techniques (**Normalization, Noise Reduction**), to improve the data quality, resulting in accurate models.

Data Engineer | ZS Associates | Pune, India (Nov 2020 - July 2022)

- Engaged with **healthcare** client, achieving $1 million cost saving through multi-team collaboration, understanding business **KPIs** and dependency skills, and applied advanced **Python, Spark, SQL** expertise to develop automated **ETL data pipelines**.
- Built scalable ETL pipelines on cloud data platform for data ingestion of 20+ million data points in **Redshift** from enterprise **APIs** and files via **Python, AWS Lambda, EventBridge, S3, Step Functions, Boto3** facilitating event-driven reporting.
- Developed and deployed microservices for **APIs** on **AWS** using **Python, Lambda, IAM, DynamoDB, CodePipeline**, and **CloudFormation**, implementing **RESTful APIs** and 2-tier microservice architecture with technical documentation.
- Analyzed complex **SQL** queries, debugging **45+ failed in-house ETL** batch jobs daily for smooth report transmission to downstream.

Data Science Intern | NextLeap Aeronautics | Pune, India (June 2020 - Sept 2020)

- Employed **Apache Beam SDK** with Python to write complex data transformation logic that was executed in **Dataflow**, ensuring our data was clean, aggregated, and ready for analysis.
- Developed custom **DAGs (Directed Acyclic Graphs)** in **Apache Airflow** to automate the end-to-end execution of data pipelines, from data collection to processing and insights generation.

RELEVANT PROJECTS

NYC Collision Data Analysis

- Performed data processing, profiling, and analysis of 5+ million records in **Alteryx** for NYC Collisions dataset, Crafted 7+ data visualization in **Tableau** and **Power BI**, revealing collision trends, and wrote **SQL queries** for data validation.
- Streamlined database architecture and end-to-end data pipeline employing **Erwin Data Modeler/Apache Airflow**, for staging, cleansing, transformation, and integration of data into a data model involving 22 dimensions and 5 facts in **MySQL**, and **PostgreSQL**

Knowledge Graph Analysis for Company Domain Transition

- Extracted JSON response from **MongoDB's** knowledge graph **API**, executing **XML** and **SQL queries** to obtain relevant domain keywords.
- Utilized **GraphQL (Neo4j)** to identify competitors through **data wrangling and NLP**, combining patent keywords with **API** results.

Credit Card Fraud Detection using Efficient Feature Engineering

- Conducted **data cleaning, variable creation, feature selection, and model exploration**, reducing model training time by 40%.
- Successfully implemented Bagging and Boosting models (**Logistic Regression, SVM, Decision Tree, Random Forest, XG Boost, CV, MLP, LGBM**) that resulted in savings of $21 million and achieved an FDR@3% of 56.98%.

Restaurant Q/A System using RAG-based LLMs

- Developed and implemented an advanced restaurant recommendation system using **Llama 2 LLM, optimizing, fine-tuning, and RAG techniques**, integrating real-world Yelp data for nuanced and contextually relevant recommendations.
- Applied a multi-step methodology involving data preprocessing, **vector embeddings** using AllMiniLM-l6v2, and **Pinecone VectorDB** for efficient storage. Conducted thorough evaluation of **Hugging Face, GPT 3.5, Llama 2, GPT 4** with emphasis on **RAG techniques**, using **BLEU** and **ROUGE** metrics for comprehensive model comparison and performance assessment.