Rahul Singh

Open to relocate, PA | singhrahulsb@gmail.com | +1 (551)-280-6143 | Portfolio | LinkedIn/in/rahuls0 | github.com/singhrahulbrijesh

Summary

Data Scientist and AI/ML Engineer with **4+** years of experience developing and deploying machine learning solutions in healthcare, robotics, insurance, and advanced analytics. Expertise in creating scalable **ML models**, optimizing processes, and delivering actionable insights. Adept at collaborating across cross-functional teams to implement AI-driven solutions for real-world impact.

Skills

Programming Language: Python, C++, R- Language, SQL, Unix, SAP.

Machine Learning & AI: Tensorflow, PyTorch, Keras, Scikit-learn, NLP, Transformers (Hugging Face, BERT), CNN, RNN, LSTM, LightGBM, CatBoost.

Statistical Methods - Linear Models, Regression, Sampling, Hypothesis Testing, A/B Testing, Forecasting, Text Mining.

Big Data & Analytics Tools: Hadoop, Spark, Kafka, Hive, Informatica, Data Stage, Tableau, Power BI, PowerPoint, SSRS, Alteryx, Qlik Sense, SAS, IBM Cognos, Salesforce, Looker.

Cloud & Technologies: AWS(EC2, EMR, S3, Redshift), Azure, Git (Version control), Docker, Kubernetes, ZenML, ML Flow, Terraform, Vertex AI.

LLMs Knowledge: Llama (3, 3.1, 3.2), GPT-4, Claude, OpenChat, UltraLM, Davini003, Falcon, Alpaca, Mistral AI, Llama-3.1-nemoton-70b-instruct, Gemini, StableLM

Database & Framework: MongoDB, Flask, Django, MYSQL, Oracle, Snowflake, POSTGRE SQL, Teradata.

Work Experience

Cotiviti — **Remote, PA**

Data Scientist II — *Feb 2025 - Present*

- Managed and processed large-scale US healthcare datasets (200M+ monthly records) from providers like **AETNA, CIGNA, and USI Insurance Services**, ensuring data accuracy and reliability.
- Designed and automated ETL workflows using **Oracle SQL, HiveQL, MySQL, SAS, R,** and **Python**, reducing data processing time by **35%** and cutting production SLAs from **10 days to 7 days**.
- Conducted root cause analysis to detect anomalies, trends, and inconsistencies in healthcare data, leading to a **20% reduction** in data quality issues and improving key performance metrics.
- Created **Tableau** and **MicroStrategy** dashboards to provide data-driven insights for Cotiviti's top clients, enabling executive teams to make strategic healthcare decisions.
- Built a centralized **knowledge repository** with automated QC checklists, data maps, and reusable scripts, streamlining workflows and reducing dependency on manual intervention.
- Led day-to-day team operations in the absence of senior leaders, coordinating with cross-functional teams and stakeholders to ensure smooth project execution and timely deliverables.
- Designed and deployed **Microsoft Fabric** solutions to streamline data integration and accelerate analytics processing across multiple healthcare datasets, reducing query execution time by **35%**.
- Led the development of an ML-driven fraud detection model using Azure Machine Learning, **LightGBM**, and **AutoML**, which achieved 94% precision in identifying fraudulent claims. The solution reduced financial losses by **20%** and improved claim processing efficiency.

PCG (Public Consulting Group) — **Harrisburg, PA**

Data Science Engineer — Aug 2023 – Jan 2025

- Spearheaded the development of AI/ML solutions for healthcare clients, utilizing **medical claims**, and **pharmacy data** to predict patient **risk** factors, improving **fraud** detection, and enhancing early intervention strategies by **35%**.
- Designed a data pipeline template for the healthcare domain to do data cleaning, pre-processing, feature engineering, model building, and validation reducing the project deployment time by **60%,** enabling faster delivery of predictive insights to stakeholders.
- Collaborated with clinical experts to analyze healthcare data, enabling better alignment of service offerings with community needs, leading to a **10%** increase in service enrollment.
- Designed and maintained CloudFormation templates (in **JSON** and **YAML**) to automate the deployment of data pipelines across public and private layers. Leveraged **AWS Step Functions** to manage **Lambda** deployments, optimizing resource allocation.

- Integrated and analyzed **EHR/EMR data** to create predictive models for patient outcomes, identifying at-risk patients early and enabling more effective intervention strategies.
- Optimized ETL processes during **ERP** and **CRM** migrations, reducing data processing time by 20% and enhancing data accuracy through validation and error-checking protocols.
- Customized reports to meet specific client data requirements, achieving **100%** accuracy on Medicaid reports and reducing reporting costs by approximately **$200K**.
- Developed comprehensive Confluence documentation and **CBT templates** to enhance knowledge transfer. Facilitated knowledge transfer (KT) sessions to ensure efficient onboarding and skill transfer across teams.
- Optimized **Snowflake** environments, cutting query response times by 30% and reducing storage costs by **25%** through schema redesign and efficient data partitioning.

Gannon University **Erie, PA**

Data Science Research Assistant Aug 2022 - May 2023

- Analyzed user-purchase behavior based on demographic and feedback data to derive actionable insights for predictive modeling.
- Created a Sparse Regularizer using KL-Divergence and Sparse Autoencoder for stable training in Cold-Start Recommendation problems and implemented GAN-based algorithms (LSRGAN, WGAN) for Zero-Shot Classification with PyTorch.
- Deployed the SRLGAN model, achieving **0.53 precision**, **0.52 NDCG**, and **0.7 MRR** in Cold-Start Recommendations, with a **Top 1 accuracy of 0.64** on the AWA dataset.

Make My Clinic Pvt ltd. **Visakhapatnam, India**

Data Scientist July 2019 - May 2021

- Led quality assessment for datasets exceeding **9M** records, detecting and addressing **150**+ anomalies across clinical studies.
- Increased data accuracy and compliance by **50%** by automating validation checks with custom **macros** in SQL and **SAS**, cutting project completion time by **2X**.
- Developed predictive models using survival analysis techniques (**Kaplan-Meier**, **Cox Proportional Hazards**) in Python, SAS, and SQL. Created reports covering treatment patterns, survival rates, and healthcare utilization for **10K+** patients, contributing to a **15%** improvement in study efficiency and decision-making support.
- Applied statistical modeling, **hypothesis testing**, and sampling theory to evaluate model performance and design experiments, contributing to data-driven decision-making and effective **A/B testing** for model optimization.

Education

Gannon University **Pennsylvania, USA**

M.S. Computer Information Science- Data Science Aug 2021 - May 2023

Mumbai University **Mumbai, India**

Bachelors - Information Technology Jun 2016 – July 2019

Research

Automating Patch set generation from code review comments using LLM - LINK **Nov 2023 – June 2024**

- Designed an automated patch set generation system using **GPT-4o** and Llama 3.2, leveraging **Qdrant DB** for vectorized storage and retrieval to improve code review efficiency. Conducted in-depth research on Apache projects, including Kafka, Spark, and Airflow, to ensure real-world relevance and scalability.
- Built retrieval-augmented generation (**RAG**) workflows by embedding large text corpora and vectorizing GitHub pull request data with Qdrant DB, enabling AI-driven code analysis with over **80%** similarity to human-generated reviews and enhancing integration with real-world open-source systems.

Common Defects in Modern Web Browsers using Knowledge Embedding in LLM - LINK **Dec 2022 – Aug 2023**

- Leveraged Selenium to scrape large datasets (**6M** from Firefox, **8M** from Chrome) and applied NLP, SQL (1,000+ queries), and GPT-4.o to analyze defects, achieving a high precision and recall rate with an F1 score of **94.63%**.
- Analyzed **370K+** Firefox and **143K+** Chromium bugs, identifying defect-prone components and high-effort issues using agile methodologies and NLP models like BERT. Improved bug-fixing prioritization by **30%**, boosting browser stability, reducing debugging time, and enhancing both user experience and developer efficiency.