Rahul Singh

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

Summary

Data Specialist with **4+** years of experience in data management, machine learning, and business intelligence within **Healthcare**, **Insurance**, and **Institutional** domains. Demonstrated expertise in deploying machine learning models, automating processes, and collaborating with stakeholders. Achieved significant improvements in data-driven decision-making, predictive analytics, and operational efficiency in highly regulated environments.

Skills

Programming Language: Python, R- Language, SQL.

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

Cloud & Technologies: AWS, Azure, GCP, Flask, Django, Git.

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

Data Science - TensorFlow, Pytorch, NLP, Transformers (Hugging Face, BERT), Linear Models, Sampling Methods, Data Mining, Neural Networks(CNN, RNN, LSTM), NumPy, LightGBM, CatBoost, GLM/Regression, Matplotlib.

Libraries & APIs: TensorFlow, Pandas, Boto3, AWS Wrangler, AWS Glue, Aws Redshift, NumPy, PySpark.

Work Experience

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

Data Specialist — Aug 2023 - Present

- Spearheaded production deployments, facilitating deployment meetings and implementing robust monitoring by integrating **Splunk** with **AWS CloudWatch**, significantly improving real-time system monitoring and issue tracking.
- 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 and improving deployment reliability.
- 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.
- Enhanced code reliability by implementing extensive unit and integration testing using pytest and Unit test libraries. Utilized Tox for library version management, accelerating deployment processes and ensuring compliance with safety standards.
- Built a generalized template for the healthcare domain to do data cleaning, pre-processing, feature engineering, model building, and validation reducing the project deployment time by **60%**.
- Customized reports to meet specific client data requirements, achieving **100%** accuracy on Medicaid reports and reducing reporting costs approximately by **$200K**.
- Responsible for taking ownership of the project, evaluating existing workflows, and recommending insights identifying key issues, increasing efficiency by **25%**.
- Performed **ETL** migration, facilitating seamless data transfer between **ERP** and **CRM** systems. Achieved a **20%** reduction in processing time and improved data quality by implementing validation and error-checking mechanisms.
- Engineered and implemented **Informatica** to streamline data integration, automate processes, and generate essential reports, improving data quality and accessibility for analysis and reporting.
- Partnered with the Business Systems team to design and implement data-driven solutions using **Tableau** and **Excel**, enabling insights-driven decision-making and trend identification.

Gannon University — **Erie, PA**

Data Science Research Assistant — Aug 2022 - May 2023

- Analyzed and tested the null hypothesis for **user-purchase behavior** derived from user demographics and purchase feedback data.
- Innovated a Sparse Regularizer for leveraging sparsity in behavior space inspired by KL-Divergence & Sparse Autoencoder to accomplish training stability and avoid seen class over-fitting for User and Item Cold-Start Recommendation problems.
- Deployed a sparse adversarial model SRLGAN for hybrid User Cold-Start Recommendation using the proposed Sparse Regularizer achieving a precision of **0.53**, Normalized Discounted Cumulative Gain score of **0.52**, and Mean Reciprocal Rank score of 0.7.
- Implemented algorithms in GANs like LSRGAN and WGAN for Zero-Shot Classification using PyTorch accomplishing a top 1 accuracy score of **0.64** on popular datasets like AWA with a training data size of approx. **40K** seen and **10K** unseen class.

Make My Clinic Pvt ltd. **India**

Data July 2019 - May 2021

- Developed and maintained scalable e-commerce platforms for marketing products, achieving an **88%** functionality and performance rating with Python and related frameworks.
- Deployed ML models on AWS (**EC2, S3, RDS, Redshift**) to ensure high availability, scalability, and data security for production environments, following best practices in MLOps for deployment, monitoring, and scalability.
- Deployed ML and NLP models using techniques such as **Logistic Regression**, **Decision Trees**, and **Time-series** analysis to solve business challenges and support predictive analytics in areas like customer behavior modeling.
- Applied statistical methods, 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.
- Established CI/CD pipelines using **Jenkins** and AWS **CodePipeline**, reducing release cycles by 30% and enhancing deployment efficiency.

Education

Gannon University GPA: 3.875 **Pennsylvania, USA**

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

Mumbai University GPA: 3.25 **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**

- Evaluated the effectiveness of pre-trained LLMs, including **GPT-4**, **GPT-3.5-turbo**, **Llama 3.2,** and others, in replicating human tasks related to code review.
- Applied research to Apache projects (Kafka, Spark, Airflow), focusing on real-world pull requests, automating outcome assessments, and integrating matched code changes with 80%+ similarity. Lower similarity matches were directed for manual review, collaborating with developers for further insights.
- Leveraged **Qdrant DB** as a vector database to store and index pull requests and patch sets, enabling efficient similarity searches and retrieval of relevant code changes.
- Utilized **Docker** to containerize processes, ensuring that pull request evaluations and patch set comparisons were run consistently and in isolation for each project. This containerized approach streamlined deployment and scaled the evaluation of different patch sets across multiple repositories.

Categorizing the common defects in modern Web browsers by Knowledge Embedding to LLM - LINK **Dec 2022 – Aug 2023**

- Performed web scraping using Selenium to collect and extract bug data from repositories and databases of Firefox and Chrome, resulting in datasets of **6 million** and **8 million** Large Datasets respectively.
- Implemented K-means clustering and Bert model to refine the dataset by grouping similar synonyms and removing less important words occurring below **0.3**%, resulting in a more accurate and focused dataset.
- Developed and deployed a GPT-4 model, achieving a **30%** boost in defect categorization accuracy and cutting bug analysis time by **40%**, enhancing overall efficiency.

Projects

Thyroid Detection:- GitHub Link **Feb 2022 - May 2022**

- Designed a scalable end-to-end machine learning pipeline utilizing **clustering** and **classification** techniques to accurately determine compensation for hypothyroidism in patients and the AUC score of **0.9%**.
- Developed models focusing on clustering and outlier detection in high-dimensional medical data, leveraging Python and deep learning frameworks like **TensorFlow** and **PyTorch**.
- Achieved an impressive accuracy of **94.5%**, ensuring reliable and accurate predictions for patient diagnosis.