# Rahul Singh

+1 551-280-6143 | Erie, PA | singhrahulsb@gmail.com | linkedin.com/in/rahul | Portfolio | github.com/rahul

## Experience

**Data Science Engineer** Aug 2023 – Present
*Public Consulting Group* *Harrisburg, PA*

- Designed AI solutions using **TensorFlow** and LLMs like **GPT** and **BERT** for text summarization, sentiment analysis, and enterprise applications, improving operational efficiency and reducing manual analysis time by **40%**.
- Developed and optimized ETL pipelines using **PySpark**, **AWS Glue**, and **Step Functions** to extract, transform, and load large-scale healthcare datasets, increasing data processing efficiency by **50%** and enabling high-quality inputs for predictive analytics and machine learning models.
- Built a pipeline for a centralized BI visualization tool to monitor various healthcare programs and intervene with the least efficient members(customers), resulting in a **20%** reduction in the effort required from team members.
- Orchestrated machine learning model deployments using **AWS SageMaker** for real-time inference, **AWS Lambda** for serverless scalability, and Step Functions for workflow automation, reducing deployment time by **25%** and enabling seamless integration into production systems.
- Engineered an ensemble model combining **Logistic Regression**, **XGBoost**, and **SVM** to achieve **92%** accuracy in predicting order processing issues; findings directed the resolution of three critical bottlenecks in customer service.

**Data Science Research Assistant** Aug 2022 - May 2023
*Gannon University* *Erie, PA*

- Analyzed predictive models using deep learning and feature engineering, achieving **95%** accuracy across organizational verticals while reducing data redundancy by **20%**.
- Implemented GAN architectures, including LSRGAN and WGAN, for zero-shot classification and recommendations, achieving top-1 accuracy of **0.64** and precision of **0.53** on datasets with **40K** seen and **10K** unseen classes.

**Data Scientist** July 2019 - May 2021
*Make My Clinic Pvt Ltd* *India*

- Led quality assessment on **9M+** clinical records, identifying **150+** anomalies and automating validation with SQL and SAS macros, improving data accuracy by **50%** and halving project time.
- Designed survival analysis models **(Kaplan-Meier, Cox)** in Python, **SAS**, and **SQL**, creating reports on treatment patterns and survival rates for **10K+ patients**, boosting study efficiency by **15%**.
- 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.

## Technical Skills

**Languages**: Python, SQL (Postgres, Snowflake), NoSQL(MongoDB, DynamoDB, Cassandra) JavaScript, R
**Big Data & Analytics**: Big Data & Analytics Tools: Hadoop, PySpark, Spark, Hive, Databricks, Informatica, Airflow, Informatica PowerCenter, Data Stage, Tableau, Power BI, SSIS, SAS
**Libraries & API**: TensorFlow, Pytorch, Boto3, Pandas, NumPy, Spark, AWS Wrangler, AWS Glue, AWS Redshift, XGBoost, OpenCV, Keras, MapReduce, Scikit-learn, NLP.
**LLMs & tools Knowledge**: Llama(2,3.1,3.2), Gpt-4o, BERT, Claude 3, PaLM 2, Davini003, Mistral AI, Gemini
**Cloud & Technologies**: AWS(EC2, S3, Lambda, Cloudfront), Azure, Git, Docker, Kubernetes, ML Flow, Splunk
**Monitoring & CI/CD**: AWS CloudWatch, Elasticsearch, Jenkins, Gitlab, CI/CD, AWS CodePipeline, Github Actions

## Education

**Gannon University** Erie, PA
*M.S. Computer Information Science, Minor in Data Science* *Aug 2021 - May 2023*

**Mumbai University** Mumbai, India
*B.S. Information Technology* *June 2016 - July 2019*

## Projects

**Automating Patch Set generation from code review comments using LLM - LINK** 2020 – Present

- Assessed the performance of pre-trained LLMs, including **GPT-4**, **GPT-3.5 turbo**, and **Llama 3.2**, by analyzing a **30K** patch set against historical human-generated data; findings optimized understanding of AI capabilities in code reviews.
- Conducted in-depth research on Apache projects, specifically **Kafka**, **Spark**, and **Airflow**; enhanced real-world pull request relevance by automating outcome assessments and integrating code changes with over **80%** similarity.

**Common defects in modern Web browsers by KE to LLM - LINK** Dec 2022 - May 2023

- Leveraged Selenium to scrape large datasets (**6M** from Firefox, **8M** from Chrome) and applied **NLP**, **SQL** (**1,000+** queries), and GPT-4o to analyze defects, enhancing insights into bug patterns and browser performance.
- Employed agile methodologies for continuous model improvement and integrated advanced NLP models like BERT to identify critical browser issues faster.