RAJAT KAPGATE

Open to Relocation | +1 (812) 553-2822 | rhkapgate@gmail.com | linkedin.com/in/rajat-kapgate | github.com/rajatk9962

PROFESSIONAL EXPERIENCE

Analyst - Strategic Patient Insights | Piedmont Healthcare, Atlanta, GA   **August 2025 – Present**

- Leveraged **SQL** to analyze provider metrics and national birth rate data, explaining maternity campaign underperformance and enabling a confident **$150K+** reinvestment decision.
- Developed a **GPT-4o** powered **NLP** pipeline using prompt-completion to classify, summarize, and **quantify 10,000+** free-text patient survey responses, enabling faster patient experience reporting across 5 departments.
- **Built** 2 interactive Tableau dashboards analyzing **1M+ office visit records** spanning **3 years** to identify **trends** and **seasonal patterns**, enabling marketing teams to optimize campaign timing and improve **engagement by 15%.**

Research Data Scientist | Indiana University School of Optometry, Bloomington, IN   **May 2024 – Present**

- Analyzed 500+ GB of infant head and eye movement data (NIH Grant EY032897) to identify early markers of eye disorders; performed ANOVA across five age groups, revealing statistically significant motor control variations.
- Preprocessed **time series data** using techniques like IQR filtering, rolling sum, low pass filtering and non-max suppression to isolate meaningful head motion segments, enhancing data quality by **40%**.
- Engineered **advanced data visualizations**, including head movement reconstruction with Unity and Open3D, KDE, polar plots, and correlation maps to analyze infant head dynamics.

Data Scientist | Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT   **May 2024 – Nov 2024**

- Built a **Retrieval-Augmented Generation (RAG)** system on proprietary organizational data using **Azure GPT-4o**, **FAISS and LangChain**, and systematically evaluated response quality with **Ragas**, reducing research effort by 40%.
- Performed **quantization-aware training** by fine-tuning a **Llama 3-8B** model with **QLoRA** on Jardiance reporting data, validated outputs with **semantic evaluation** and SME review, integrating into a reporting assistant that cut reporting turnaround by 70%.
- Implemented a **FinOps** cost analytics dashboard with **Streamlit**, integrating cloud cost data for better visibility. Monitored Jenkins CI/CD pipelines and performed **root cause analysis**, uncovering inefficiencies and cutting compute costs by **$200K+.**
- Orchestrated an **ETL pipeline** to process **2M+** drug price records from the Nuro API, writing analytic **SQL** with window functions in Amazon Redshift, scheduling jobs with cron, and storing curated outputs in AWS S3.

Data Scientist | TCS Research, Mumbai, MH   **Jun 2021 – Aug 2023**

- Designed a multi-stream **CNN–LSTM** sign language recognition model in **PyTorch** using Mediapipe pose key points, reducing per-frame input size from **200K+ RGB pixel values to 478 pose features** for efficient CPU deployment.
- Achieved **24.4% WER** on the RWTH-PHOENIX benchmark with this compact setup, outperforming standard models while remaining suitable for **real-time inference**.
- Automated team expense reporting with Python, cutting manual work by **40%** and improving accuracy, and added **ARIMA** forecasting in **statsmodels** to predict future spend.
- Served as a key resource for **AI** and **machine learning**, mentoring six associates and demystifying model workflows for non-technical stakeholders, resulting in a 20% boost in team efficiency and informed decision-making.

EDUCATION

Indiana University Bloomington, USA   **Aug 2023 – May 2025**
Master of Science in Data Science   **GPA: 3.94/4.00**
Coursework: Data Mining, Machine Learning, Advanced Database Concepts, Statistics, Data Visualization, Algorithms

University of Mumbai, India   **Aug 2017 – Jun 2021**
Bachelor of Engineering in Computer Engineering   **GPA: 3.50/4.00**
Coursework: Big Data Analytics, Advanced DB, Elements of Artificial Intelligence, Data Structures, Exploratory Data Analysis

TECHNICAL SKILLS

- **Programming Languages:** Python, R, SQL
- **Business Intelligence:** Tableau, Power BI, Excel, Looker Studio, GeoPandas, Seaborn, NumPy, Pandas
- **Frameworks and Tools:** Apache Spark, NLTK, Ragas, Langchain, LlamaIndex, PyTorch, Tensorflow
- **Databases:** MySQL, BigQuery, PostgreSQL, MongoDB
- **Cloud Tools:** Azure (Databricks, Data Factory), AWS (S3, Redshift, Glue, Lambda, Sagemaker), GCP (BigQuery), Kubernetes
- **Project Management:** Confluence, Jira, Agile, Waterfall, SDLC, Kanban, CRM, ERP

ACADEMIC PROJECTS

Photo Realistic Face Generation using Generative Adversarial Networks

- Devised a controllable GAN model to generate high-resolution human images with options to select from over 10 facial features. Incorporated the Wasserstein Loss function and an auxiliary 2D CNN to unravel the latent space and estimate feature distributions.

Vio-Later: Predictive Prevention of Traffic Violations and Road Accidents

- Developed a traffic violation prevention system using New York traffic data, leveraging **GeoPandas** for city map visualization using **ensemble** of **XGBoost, ANN, Decision Trees** to predict high-risk violations and reduce accidents.

Patient Outcomes (Tableau)

- Unearthed **5 key trends** in MIMIC-III dataset through **data visualization**, providing **actionable insights** into patient outcomes.