RAJENDRA KUMAR

+1 (812) 803-4330 ◇ kummrajnn@gmail.com ◇ linkedin ◇ GitHub ◇ Portfolio

EDUCATION

Master of Science in Data Science (Big Data Systems) — Aug 2023 - May 2025
Indiana University Bloomington — GPA: 3.7

PG-Diploma in Big Data Analytics & Machine Learning — Feb 2018 - Aug 2018
Centre for Development of Advanced Computing — Grade: A

Bachelor of Engineering in Computer Science & Engineering — Jun 2013 - Jul 2017
Rajiv Gandhi Proudyogiki Vishwavidyalaya — GPA: 8.02

WORK EXPERIENCE

Lead Data Scientist (AI/ML) - Heartland Network — June 2025 - current

- Engineered a document processing pipeline with LangGraph agents, reducing researcher literature review time by 75%.
- Devised an intelligent RAG system, improving document discovery relevance by 90% with sub-second responses.
- Developed a scalable vector DB with ChromaDB, processing 2000+ papers into 150k+ searchable chunks embeddings.
- Built RAG system embedding 1k+ property documents with Zapier AI/n8n automation, cutting response time by 60%.

Research Assistant (Gen AI) - Indiana University Bloomington — Sept 2024 - May 2025

- Designed and built full-stack web app with React, Flask, and PostgreSQL, scaling for 15+ users at once via FastAPI.
- Trained Generative Origami models using Stable Diffusion and GAN, generating origami images from custom datasets.
- Fine-tuned SD 1.5 using LoRA/QLoRA for Origami generation; monitored metrics/parameters for 5+ models via W&B.
- Developed RAG pipeline for Origami AI, indexing 5K+ patterns and integrating GPT for real-time folding instructions.

Senior Data Scientist - Target Corporation — Apr 2023 - Jun 2023

- Built end-to-end credit fraud detection pipeline with EDA, feature engineering, preprocessing, class imbalance handling.
- Achieved 0.97 ROC AUC using a Random Forest model, delivering 0.99 precision and 0.80 recall on the fraud class.
- Created real-time analytics dashboards on Domo & Greenfield, offering key insights to management for decision-making.
- Engineered SQL/Python ETL pipelines using CTEs, procedures, delivering 40% faster retail-financial data processing.
- Built XGBoost model classifying 250+ retail categories from unstructured reviews, raising fiscal insight accuracy 79%.

Lead Data Scientist - Sutherland Global — Sep 2018 - Mar 2023

- Led the project Propensity-To-Pay (P-T-P), achieved a 70% increase in response time, and scaled it to multiple clients.
- Developed time-series forecasting model predicting claims flow and conversion rates, optimizing workforce by 30%.
- Applied sampling theory, statistical methods and calibration techniques in model development and claims forecasting.
- Engineered EDA and scalable data transformation pipelines with PySpark on Azure, generating advanced analytics.
- Auto-tuned HiveQL and SparkSQL workflows, cutting query latency by 30% and slashing memory footprint by 50%.
- Developed Smart-Doc from scratch, using Python, JavaScript, and OCR to process EHR/PHI unstructured documents.
- Orchestrated and monitored parameters, metrics, and artifacts using MLflow and Azure Databricks for 20+ ML models.
- Engineered data pipelines to process files like XLSX, CSV, TXT, PDF, and image, converting them into HL7 format.
- Developed the web scraper tool using Python (BeautifulSoup) and shell script to extract census data from 50k URLs.

SKILLS

Certifications	**AWS ML Associate**, **AWS AI Practitioner**, **Azure AI**, **Databricks Gen AI**
Programming Languages	Python, SQL/NoSQL, Scala, Java, R, JavaScript, Shell scripting
Web & Databases	HTML, CSS, JavaScript, Flask, Vector DB, SSIS, PostgreSQL, MongoDB, Neo4j
Libraries & Frameworks	Pandas, PySpark, NumPy, Sklearn, TensorFlow, PyTorch, FastAPI, Airflow, CI/CD
Analytics Tools	AWS, GCP, BigQuery, Bedrock, RStudio, Tableau, Power BI, Hypothesis, A/B testing
ML/DL algorithms	Random Forest, XGBoost, RNN, Panel data modeling, LLM, Gen AI, GPT, GNN, ARIMA, LangFuse, LangGraph, CrewAI, MLflow, MCP, Docker, DBT, W&B, LSTM

PROJECTS

Propensity-To-Pay (Predictive Modeling):- *Tech Used: Python, SQL, PL/SQL, Hive, Random Forest, XGBoost.* **Agile**

- Devised PySpark scripts to process transaction data into account-level aggregates, storing in Cloudera Hive tables.
- Developed P-T-P model from scratch and achieved an accuracy of 93% by hyperparameter tuning using GridSearchCV.
- Enhanced ensemble model (95% accuracy) to predict insurance actions/status codes, increasing client revenue by 13%.

Clinical Text - Knowledge Extraction and Analysis (NLP):- *Tech Used: Python, SQL, Java, MongoDB, UMLS.* **Agile**

- Wrote Python script to process high-volume unstructured clinical text corpora, retrieving /storing data from MongoDB.
- Integrated Hadoop HDFS and Shell scripting for the data flow design in 3 layers: Staging, Gold, and Datamart layers.
- Used spaCy, NLTK, and BERT for NLP tasks, assigning ICD/CPT/SNOMED codes to text corpora with 88% accuracy.

Pneumonia Detection - Chest X-Ray Images:- *Tech Used: Python, CNN, PyTorch, AlexNet, ViT, CUDA*

- Processed 5,856 chest X-rays using resizing, normalization, and channel conversion for pneumonia detection models.
- Implemented grayscale/3-channel conversion and multi-GPU training (8x NVIDIA A40) for optimized DL workflows.
- Compared ResNet18(79.97%), AlexNet(78.85%), CNN(77.56%), and ViT(63.46%) in multi architecture analysis.