Dhruvraj Singh Rathore

(737)206-1179 | dhruvrajrathore2011@gmail.com | Linkedin | Github

Education

Master of Science in Data Science — Aug. 2024 – Dec. 2025
Texas A&M University — *College Station, TX*

Bachelor of Technology in Computer Science Engineering — Jul. 2018 – May 2022
SRM Institute of Science and Technology — *Chennai, India*

Technical Skills

Programming & Data Science: Python, SQL, R, SAS, Pandas, NumPy, Matplotlib, Scikit-learn, Shell Script
Databases & Cloud Computing: MySQL, DynamoDB, Snowflake, AWS Suite (EMR, S3, EC2, Glue, Lambda)
Big Data & Machine Learning: Spark, PySpark, TensorFlow, PyTorch, LLMs (Large Language Models), BERT, Ollama, LangChain, Statistical Modeling, Hypothesis Testing
Tools & Platforms: Git/GitHub, CI/CD, Power BI, Tableau, Data Build Tool (DBT), Apache Airflow, Docker

Experience

Student Research Assistant — Oct. 2024 – Present
School of Public Health, Texas A&M University — *College Station, TX*

- Conducted **descriptive statistical analysis** on Medicaid data to derive actionable insights using Python and SAS.
- Streamlined graph generation for project presentations using R, **reducing processing time by 30%.**
- Performed **hypothesis testing using T-tests and Pearson correlation** to identify relationships between target columns, and **utilized Q-Q plots** to analyze distribution similarities across datasets over multiple years.

Data Analyst — Dec. 2022 – Jun. 2024
Draup Business Solutions — *Bangalore, India*

- **Led cross-functional initiative** for the development of a **data quality monitoring system**, leveraging **Airflow** to process 200M+ records daily, resulting in 40% improved data accuracy.
- Engineered robust ETL pipelines using **PySpark on AWS EMR**, optimizing data ingestion workflows and **reducing end-to-end processing time by 45%** while maintaining data integrity.
- Integrated a serverless data retrieval system using **AWS Lambda, S3, and DynamoDB**, streamlining ad-hoc client data requests and **reducing response time by 60%**.

Data Scientist — Aug. 2021 – Nov. 2022
HighRadius Corporation — *Hyderabad, India*

- Automated data extraction and preprocessing using **Python and SQL**, reducing analysis time by 50%.
- Built **Machine Learning models using LightGBM and Random Forest** to predict customer payment dates, improving cash flow forecasting accuracy by 40%.
- Continuously monitored and **evaluated the performance of deployed models for 50+ accounts**, updating or refining them as needed alongside senior analysts.
- Created Power BI dashboards to visualize predictions, effectively communicating insights to non-technical stakeholders.

Projects

Personalized Academic Research Assistant | *NLP, RAG, Langchain, LLM* — Dec. 2024 – Jan. 2025

- Built an Academic Research Assistant using RAG and LangChain to retrieve, rank, and summarize papers.
- Used FAISS with SciBERT embeddings for retrieval and fine-tuned BERT for ranking.
- Integrated Ollama3.2 for summarization and multi-turn conversational queries.

Cotton Field Detector | *Python, CNN, U-NET, Pytorch, Computer Vision* — Nov. 2024 – Dec. 2024

- Developed an automated method to identify and map cotton crop areas from satellite imagery of the United States using UNET algorithms in PyTorch.
- Performed segmented image classification to isolate cotton crops from other vegetation.
- Calculated total cotton acreage by analyzing pixel coverage of masked areas, providing accurate crop area measurements.

Metro Interstate Traffic Volume | *Machine Learning, Statistical Techniques* — Oct. 2024 – Nov. 2024

- Built a traffic congestion model using scikit-learn, with feature scaling, one-hot encoding, and time-series analysis.
- Applied Random Forest, Lasso, Ridge, Linear, and Polynomial Regression with cross-validation.
- Tuned models with GridSearchCV, and found Polynomial Regression most accurate with RMSE and R-squared.