RIJIN STALIN
Data Scientist
+1 (267)-341-9020 | rijinstalin365@gmail.com | www.linkedin.com/in/rijin-s-83455326b

Professional Summary:

- **Data Scientist with ~6 years** across data science, data engineering, and analytics.
- Deliver applied ML in **fraud, recommendations, forecasting, and operations**, aligned to clear product/ops KPIs.
- End-to-end workflow: **problem framing, data prep, modeling (scikit-learn/PyTorch), A/B testing, deployment (FastAPI/Docker), and monitoring (MLflow/Evidently)**.
- Build reliable pipelines on **Databricks (PySpark/Delta/Unity Catalog)** and **Snowflake** with **dbt** for transformations and documentation.
- Implement **data quality and governance** using **Great Expectations**, data contracts, and basic lineage/observability.
- Use **SQL/Python** daily; produce executive-ready insights with **Power BI (DAX)** and **Tableau**.
- Practice **CI/CD** with **GitHub Actions** and versioned artifacts; maintain runbooks and on-call procedures to meet SLOs.
- Experience in **e-commerce/transportation logistics** (Amazon) and **financial services/risk** (GTE).
- Work cross-functionally with Product, Engineering, and Operations; write clear documentation and mentor junior teammates.
- Primary focus as a Data Scientist; also experienced contributing as a Data Engineer or Senior Data Analyst.
- Developed predictive credit scoring and loss forecasting models across **installment loans, revolving credit lines, and BNPL portfolios**.

Tools and Technologies:

Programming & Analytics: Python (Pandas, NumPy, scikit-learn, PySpark, Spark(scala), SQL (advanced; window functions, CTEs), Jupyter

ML & Experimentation: Supervised/unsupervised learning, time-series forecasting, feature engineering, experiment design/A/B testing; MLflow (tracking/registry), Evidently (drift/quality)

Data Engineering & Lakehouse: Databricks (PySpark, Delta, Unity Catalog), Snowflake (Tasks, Streams), dbt (models/tests/docs/exposures), Airflow/Prefect, Kafka/Kinesis, CDC (Debezium – familiar)

Data Quality & Governance: Great Expectations, data contracts, Open Lineage; Catalogs: Alation/Collibra (familiar)

MLOps & Serving: FastAPI, Docker, CI/CD (GitHub Actions/Azure DevOps), monitoring/observability (Datadog)

BI & Visualization: Tableau, Power BI (DAX, Power Query), Looker (LookML)

Cloud Platforms: AWS (S3, Lambda, Redshift, EMR, SageMaker basics), GCP (BigQuery, Pub/Sub), Azure (Synapse/ADF basics)

Education:

Master's in Business Analytics, **St. Francis College**

Bachelor's in Electronics and Communication Engineering

Certifications:

- **Databricks Certified Data Engineer Associate**
- **Google Cloud Professional Data Engineer**
- **Microsoft Certified: Azure Data Scientist Associate**
- **Snowflake SnowPro Core Certification**

Professional Experience:

GTE, FL, USA **Jan 2024 – Present**

Data Scientist

Responsibilities:

- Designed and deployed **credit risk and fraud detection models** using **Python and SQL** to support **loan approvals, refinance strategies, and credit line increases (CLI)** in a regulated financial environment.
- Built scalable **feature engineering pipelines** capturing borrower behavior, transaction patterns, repayment history, and risk signals using **PySpark and Databricks**.
- Developed end-to-end **financial data pipelines** ingesting transactional, customer, and third-party risk data into **Snowflake**, ensuring accuracy, consistency, and audit readiness.
- Implemented **data validation, reconciliation, and quality checks** to ensure regulatory-grade reliability of credit and fraud datasets used for analytics and modeling.
- Partnered with **risk, compliance, and product teams** to translate business and regulatory requirements into analytical solutions aligned with internal controls and governance standards.
- Migrated legacy **SAS and Excel-based financial workflows** to automated **Python and Spark pipelines**, reducing manual effort and operational risk.
- Delivered **executive-level reporting and dashboards** highlighting credit performance, fraud trends, approval rates, and portfolio risk metrics.
- Documented **data lineage, metric definitions, and model assumptions** to support governance, compliance reviews, and stakeholder transparency.
- Architected and built **GenAI and RAG frameworks from scratch**, enabling intelligent document search, decision support, and analytics augmentation using LLMs.
- Implemented **vector search pipelines** using **FAISS and Pinecone**, including embedding generation, indexing, retrieval, reranking, and response post-processing.
- Designed and deployed **FastAPI-based inference services with JWT authentication** to expose GenAI and ML capabilities via secure APIs.
- Developed large-scale **feature and retrieval pipelines using PySpark and Scala**, supporting both ML training and real-time GenAI inference workflows.
- Owned the **ML and GenAI lifecycle using MLflow**, including training, fine-tuning, versioning, deployment, monitoring, and automated retraining.
- Integrated **batch and streaming data pipelines** using **Kafka and Databricks Auto Loader** to support real-time risk signals and AI-driven insights.
- Implemented **CI/CD pipelines and containerized deployments** using **Docker, Kubernetes, GitHub Actions, and Terraform**, ensuring scalable and reliable production systems.

Project: GenAI-Powered Credit Risk & Fraud Intelligence Platform

- Built an **end-to-end GenAI platform** combining **credit risk analytics, fraud detection, and LLM-based insights** for financial decision-making.
- Developed **RAG pipelines** to retrieve and summarize credit policies, transaction histories, and compliance documentation using **vector databases**.
- Created **secure FastAPI inference services** for real-time fraud scoring and AI-assisted risk explanations.
- Engineered **PySpark and Scala pipelines** to process large-scale transactional data for both ML models and GenAI retrieval.
- Implemented **MLflow-based lifecycle management** for model training, deployment, monitoring, and retraining.
- Enabled **executive dashboards** integrating predictive metrics with GenAI-generated insights to support faster and more accurate credit decisions.

Amazon Development Center, Hyderabad, India **Sep 2018 – Aug 2022**

Data Analyst

Project: Transportation Analytics & Exception Management Platform — Built analytics pipelines and dashboards to optimize middle-mile network performance, reduce SLA delays, and improve delivery efficiency across regional operations.

Responsibilities:

- Analyzed **middle-mile and last-mile logistics** data using **SQL, Python (Pandas)**, and **Redshift**, improving route efficiency and reducing delivery cost per shipment by **10–12 %**.
- Automated **data ingestion and transformation pipelines** with **Airflow** and **Python scripts**, reducing manual reporting effort by **~40 %**.
- Partnered with Data Engineers to migrate reporting to **Snowflake** and designed reusable data models for on-time performance and SLA tracking.
- Developed and maintained **Tableau dashboards** to visualize network KPIs—trailer utilization, dwell time, lane performance—used by 100+ ops leaders.
- Implemented **Data** validations on key logistics datasets, improving accuracy and timeliness of KPI reporting.
- Built analytical models to forecast **daily shipment volumes and line-haul demand**, increasing planning accuracy by **~15%**.
- Conducted root-cause analysis on SLA misses using **SQL joins, CTEs, and window functions**, identifying key factors that led to process improvements worth **multi-million-dollar savings**.
- Designed a **real-time exception-tracking dashboard** integrating data from **AWS S3, Redshift, and QuickSight**, reducing response time to incidents by **30 %**.
- Engineered and fine-tuned large language models using **prompt optimization techniques (zero-shot, few-shot, chain-of-thought)** to improve factual accuracy and contextual reasoning in business-critical responses.
- Collaborated with product and operations teams to define **data quality standards**, create runbooks, and ensure consistent KPI definitions across regions.
- Mentored junior analysts on **SQL optimization, Tableau modeling, and business storytelling**, raising analytical maturity within the transportation analytics team.

Syneos Health, Bengaluru, India **Nov 2017 – May 2018**

Data Analyst/ Intern

Project: Clinical Trial Data Quality & Performance Dashboard — Automated ETL pipelines and dashboards to track site performance, enrollment trends, and data integrity for faster regulatory reporting.

Responsibilities:

- Collected, cleansed, and analyzed **clinical and operational datasets** from multiple EHR and trial management systems using **SQL** and **Python (Pandas)**.

- ➢ Designed and automated **ETL pipelines** to process patient, protocol, and site data—improving refresh frequency from weekly to daily.
- ➢ Built and maintained **Power BI dashboards** for study progress, site performance, and data quality metrics—reducing manual reporting effort by **40%**.
- ➢ Implemented **data validation scripts** and QA routines to detect missing or inconsistent records, improving data integrity by **~25%**.
- ➢ Partnered with biostatisticians and clinical programmers to streamline data extracts for statistical analysis and regulatory submissions.
- ➢ Optimized **SQL stored procedures and joins**, reducing query runtime by **~35%**, and improving report delivery SLAs.
- ➢ Developed standardized templates and metrics for **clinical trial data visualization** (enrollment trends, adverse events, protocol deviations).
- ➢ Supported integration of **Excel VBA macros and Power Query** for ad-hoc clinical operations reporting, enabling faster data review cycles.
- ➢ Ensured compliance with **HIPAA and GCP** standards while handling sensitive patient and trial data.
- ➢ Documented ETL workflows, validation logic, and data dictionaries to improve traceability and onboarding for new analysts.