Ajay Sreekumar

ajaysreekumar.nmims@gmail.com | (775) 830-6747 | linkedin.com/in/ajay-sreekumar-nmims | github.com/AjaySreekumar47 | ajaysreekumar47.github.io

EDUCATION

University of Arizona *August 2023 - May 2025*

Master of Science, Data Science

- **GPA**: 4.0 (Cumulative); *Honors Scholar*

NMIMS University *August 2014 - May 2018*

Bachelor of Engineering, Electronics and Communications Engineering

EXPERIENCE

University of Arizona (NASA JPL & BAE Systems) — *Machine Learning Engineer* (Tucson, AZ) *Jan 2025 - Present*

- Engineered an **AI-powered NLP pipeline using LangChain framework and custom-tuned LLMs** (7B-13B parameters) with context window optimization, improving code generation accuracy by 90%
- Developed advanced **RAG architecture** integrating Tiktoken BPE tokenizer and L5-large (768-dimensional) Sentence Transformer embeddings with cosine similarity thresholding, achieving 60% improvement in domain-specific information retrieval
- Implemented data preprocessing techniques including token normalization, chunking strategies, and n-gram feature extraction, **reducing inference latency from 3.8s to 0.74s** on production workloads
- Created custom data validation framework with Pydantic schema enforcement, integrating EBNF grammar constraints converted to Lark parsers, ensuring 100% syntactic compliance with specifications

Los Alamos National Laboratory — *Data Science Intern - LLMOps* (Los Alamos, NM) *May 2024 - August 2024*

- Designed a comprehensive ML evaluation framework for comparing **12 open-source LLM (7B-34B parameters)** performance in code translation operations, using stratified sampling techniques and CodeBLEU metrics
- Published co-author for award-winning paper "**LLM-Assisted Translation of Legacy FORTRAN Code to C++**" at AISD NAACL 2025
- Engineered an **agentic feedback system** integrating compiler error diagnostics with embedding-based similarity scores, implementing a reinforcement learning loop that improved translation accuracy by 42% over baseline models
- Created multi-dimensional visualization suite using dimensionality reduction techniques (t-SNE, UMAP) to identify LLM error patterns across 5,000+ code samples, revealing statistically significant performance correlations

Accenture — *Application Development Analyst* (Maharashtra, India) *June 2018 – June 2021*

- Developed NLP-based analytics pipeline using **Python, NLTK, and SQL** with multi-table joins across 50+ CRM and B2B sales tables, uncovering actionable trends that improved lead conversion rates by 35%
- Implemented supervised and unsupervised learning models on customer transaction data, creating predictive solutions with Random Forest and SVM algorithms that increased operational efficiency by 60%
- Architected real-time data streaming infrastructure using **Apache Spark (PySpark, Spark SQL, Spark Streaming)** integrated with **Azure Event Hubs** and **Kafka**, implementing custom DataFrame transformations and optimized shuffle operations that enhanced data processing throughput from **25MB/s** to **45MB/s**
- Engineered interactive business intelligence dashboards using Tableau and AWS QuickSight with direct SQL connections to BigQuery, enabling data-driven decision making that reduced report generation time from 48 hours to 4 hours

SKILLS

Python (Pandas, PySpark) | SQL | Java | TensorFlow | PyTorch | LangChain | Generative AI | LLMs | RAG | Machine Learning | NLP | AWS | Azure | Kafka | Spark | Airflow | Terraform | CI/CD

PROJECTS

Walmart Demand Prediction (Retail Analytics) – Demand Prediction System *March 2025*

- Engineered an advanced forecasting system using **ARIMA (p=2,d=1,q=1), Prophet with custom seasonality parameters, and gradient-boosted ensemble methods**, achieving 89% accuracy on 5,000+ SKUs across multiple categories. Implemented feature engineering pipeline extracting 42 time-based variables including Fourier transformations of seasonal components, reducing forecast error (MAPE) from 23% to 11%.

Financial Services (Fraud Detection Framework) — NYC Taxi Anomaly Detection *February 2025*

- Developed a hybrid unsupervised anomaly detection system combining **isolation forests, LSTM autoencoders, and statistical thresholding** (3σ) achieving 93% precision and 91% recall on imbalanced transaction datasets. Created end-to-end ETL pipeline in **Python** processing 50GB+ of historical taxi data, implementing sliding window analysis with dynamic thresholding that identified fraudulent activities in real-time. Deployed solution using **AWS Lambda, S3, and CloudWatch** for continuous monitoring and notification

Credit Risk Analysis — (Japanese Credit Screening Project) *March 2025*

Developed a comprehensive credit risk assessment system implementing statistical and machine learning models **(logistic regression, random forests, gradient boosting, SVM)** achieving **85%+** accuracy on the Japanese Credit Screening dataset. Created an end-to-end data pipeline in Python for **preprocessing, feature engineering, and model evaluation**, with advanced **visualization of decision boundaries and feature importance**. Implemented rigorous **cross-validation** and **hyperparameter optimization** to identify **key risk factors in credit applications**. Deployed **production-ready solution** with automated preprocessing, **real-time prediction capabilities**, and model monitoring, reducing manual review time by **40%** while maintaining high-precision risk assessment.