Arun Balaji R

arun2000.balaji@gmail.com | +91 8754243305 | linkedin.com/in/arun-balaji

Summary

Lead Data Scientist with 4+ years of experience delivering analytics and AI solutions for Fortune 50 enterprises across finance, commercial operations, and marketing. Skilled in statistical modeling, demand forecasting, and scalable data engineering, with proven success in building global dashboards, automated pipelines, and enterprise RAG/GPT systems. Adept at translating complex data into actionable insights that drive efficiency, accuracy, and measurable business growth.

Skills

Programming & Data Processing: SQL, Python, R, PySpark, Polars, Pandas, NumPy

Machine Learning & Statistical Modeling: Linear & Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), k-means, Hierarchical Clustering, Hierarchical Voronoi Tessellation (HVT), ARIMA/SARIMA

Frameworks & Libraries: Scikit-Learn, LangChain, StatsModels, NLTK, spaCy, MatPlotlib, Seaborn, Plotly

AI & Advanced Analytics: Retrieval-Augmented Generation (RAG), Prompt Engineering, , Amazon Bedrock, Vector Databases (OpenSearch), Databricks (MLflow, Unity Catalog, Model Serving, Feature Store)

Version Control & Visualization: Git, QlikSense

Workflow Orchestration: Kedro, Airflow, Argo Workflows

Cloud (AWS): EC2, S3, Lambda, RDS, DynamoDB, Aurora, IAM

Experience

Apprentice Leader at Mu Sigma(Lead Decision Scientist)

GenAI Chatbot for Enterprise Insights **Jan 2025 - Present**

- **Co-led an 8-member cross-functional team** to build and deliver an Agentic Generative AI chatbot (valued at ∼$350K) for a Fortune 50 pharmaceutical company, enabling HR and business leaders to access insights via natural language queries
- Designed a **Question Definer & Retrieval flow** to classify queries into KPI, non-KPI, and ambiguous types, mapping them to the correct metrics, tables, and base SQL templates through vector search
- Integrated an automated **SQL generation and LLM summarization module**, reducing HR and workforce analytics turnaround times from 1–2 days to under 10 minutes, and cutting reliance on technical teams by ∼50%
- Enhanced enterprise decision-making speed by ∼35% across HR and commercial operations, empowering stakeholders with faster, clearer, and executive-ready insights

Conversational AI Platform with RAG **Aug 2024 - Jan 2025**

- Co-led a team of 10 in designing and deploying a custom GPT solution for a Fortune 500 airline, leveraging a Retrieval-Augmented Generation (RAG) approach with Amazon Bedrock services
- **Implemented session management and an activity monitoring plugin** to track user interactions, ensuring compliance and optimizing user experience
- Guided the product from **MVP to production with multiple enhancements**, scaling capacity from ∼1K to ∼25K concurrent users while maintaining stability and performance

KOL Profiling & Tracking **Jan 2024 - Dec 2024**

- Led a 4-member team for a Fortune 50 commercial operations group to design and deliver more than 4 interactive physician engagement dashboards integrating **8+ data sources, scaled across US, EMEA, and APAC**
- Applied **k-means clustering and weight-based percentile ranking** to segment physicians and generate performance scores, enabling objective comparisons and improving targeting accuracy
- Built a centralized physician universe with 22 custom KPIs to track popularity, expertise, and procedural volumes, streamlining engagement planning and **increasing campaign reach by ∼29%**

Demand Forecasting **May 2023 - Dec 2023**

- Streamlined ETL workflows for multi-country, multi-SKU forecasting for a Fortune 50 global pharma finance team, **reducing monthly preparation time by ∼33% and improving data completeness and accuracy by ∼14%**
- Designed SKU-level allocation and currency conversion logic, enabling 53 country teams to directly tie forecasts to operational and financial plans
- Automated classification and QA for millions of monthly records, resolving ∼87% of data issues pre-publication and elevating forecast reliability

Campaign Effectiveness & Omnichannel Marketing Engine **Oct 2022 - April 2023**

- Assessed campaign performance for a Fortune 50 commercial marketing team by applying **logistic regression** to optimize message timing, and refine content relevance – improving HCP-targeted touchpoints
- Built and scaled a data-driven omnichannel marketing engine **to identify the right patients, channels, and content at optimal times**, enhancing campaign impact
- Transitioned from proprietary systems to open-source platforms – saving ∼$480K annually, increasing transparency, and scaling across 12 brands, contributing to ∼11% sales growth

HVT Market State Analysis **March 2022 - Sep 2022**

- Applied **Hierarchical Voronoi Tessellation (HVT) with AGNES clustering** to segment complex financial metrics into interpretable market states across multiple levels.
- Enhanced anomaly detection by integrating **MAD-based thresholds** with temporal tracking, surfacing early warning signals for market shifts.
- Delivered a **visual insights framework** (state transition maps, heatmaps, summaries) that helped analysts and business teams spot risks and opportunities earlier, improving responsiveness.

Reusable EDA Solution using PySpark **Aug 2021 - Feb 2022**

- Built a **company-wide reusable framework** for Exploratory Data Analysis (EDA) on large datasets using PySpark, making it scalable and accessible across the organization (∼2,200 employees).
- Designed **plug-and-play modules** for univariate and bivariate analysis, enabling analysts of varying skill levels to quickly explore datasets without repetitive coding.
- Benchmarked performance on real-world datasets, showing a ∼**40% reduction in processing time compared to Python-based workflows**, and drove broad adoption across multiple functions.

Education

VIT Vellore, B.Tech in Computer Science (Specialized in Information Security) 2017-2021

Awards & Achievements

- Four-time Spot Award recipient (2022–2025)in recognition of performance and value delivered
- Secured 2nd Place in Overall Micro Class at SAE Aero Design East (2020) Florida, USA
- Finalist at Amrita InCTF (Capture the Flag) 2020, a national-level cybersecurity competition