# Sarath Tharayil Sreenivasan

Sheffield | P: +44 7533886825 | imsaraththarayil@gmail.com | linkedin.com/in/SarathTharayil | saraththarayil.com

## SUMMARY

**Data Scientist** with **5+ years'** experience delivering **end-to-end advanced analytics** and **machine learning solutions** within Fortune 500, multi-national organizations across the Middle East, Latin America, and North America. Proficient in generating actionable insights from large structured and unstructured datasets, developing reusable ML components, and deploying models into production environments using **modern MLOps practices**. Skilled in **Python**, **R**, **SQL**, **TensorFlow**, **PySpark**, and big data ecosystems including **Spark** and **Hadoop**. Strong stakeholder collaboration ability, experienced working in agile delivery environments to translate analytical findings into commercial and operational value. A highly collaborative **team player** who excels at translating complex technical solutions into actionable business strategies, engaging stakeholders through clear communication and compelling presentations.

## WORK EXPERIENCE

GOOGLE

**Quality Analyst** | October 2025 - Present

- Evaluate and score LLM-generated code for **correctness**, **logic**, **efficiency**, and adherence to problem requirements.
- Compare outputs across model versions and document failure cases to **support model tuning** and improvement.
- Verify code execution results, **detect edge-case errors**, and assess reasoning steps for clarity and accuracy.
- Maintain **high evaluation accuracy** and **consistency** against defined quality guidelines.

UPWORK

**Data Scientist** | January 2024 - Present

- Delivered **data science consulting projects** for global clients across E-commerce, IoT, CPG sectors.
- Specialized in end-to-end **ML pipelines**, **time series forecasting**, **LLM-powered** applications, **RAG chatbots**, **MLOps**, and **analytics dashboards**.
- Leveraged cloud platforms (**AWS, Azure**), ML libraries (**XGBoost, PyTorch, LangChain**), and data tools (**Airflow, Tableau, Snowflake**) to build scalable, impactful solutions.
- Rated **highly** across engagements for **technical depth**, **responsiveness**, and **business impact**.

MU SIGMA

**Senior Data Scientist** | September 2020 – September 2022

- Collaborated with **cross-functional agile teams** across North America, Latin America and Middle East to **develop and deploy end-to-end machine learning pipelines** for leading CPG and chemical manufacturers, delivering iterative value through sprint-based releases and continuous stakeholder feedback.
- Created multiple production-grade MLOps pipelines using **AWS S3**, **Docker**, and **GitHub workflows**, enabling CI/CD for **energy market forecasting models** with automated testing, deployment, and monitoring.
- Programmed a robust ETL orchestration framework using **Airflow DAGs** that processed **1,000+ daily API calls** from global energy markets, reducing data processing time by 18%.
- Implemented infrastructure-as-code with **Terraform** to provision and manage AWS resources, deploying containerized workflows to **ECR/ECS** reducing manual configuration time by 40%.
- Revamped data masking techniques using **Snowflake** and **Presidio** to ensure regulatory compliance by 25% while maintaining analytical utility of sensitive data.
- **Led the development** of a scalable data engineering pipeline to automate real-time data ingestion from global energy and commodity markets, orchestrating thousands of API calls daily.
- Automated data transformation using PySpark and Apache Hive, and orchestrated workflows with Apache Airflow for efficient scheduling and monitoring.
- Engineered **ensemble models** integrating **XGBoost** and **ARIMA** for robust **price and demand forecasting**, for the Saudi based client, improving forecast accuracy by 18%. Delivered insights through interactive **Tableau** dashboards.
- Built a machine learning pipeline using Python and PySpark to **identify root causes** of **supply chain disruptions**, reducing daily product shortages by **38%**, preventing losses worth $1.5M per day.
- Developed a **Market-Mix Model (MMM)** to **optimize marketing spend**s across multiple channels for a Latin American CPG client, leading to a **16%** increase in ROI in the Mexican market.
- Fine-tuned advanced statistical models to quantify **media effectiveness** across digital and traditional channels, creating a data-driven allocation framework that improved **campaign performance** by **22%**.
- Deployed **an interactive RShiny simulator** for Market-Mix Modeling, empowering business users to test and optimize various spending strategies, directly contributing to a **16%** increase in incremental revenue.

## EDUCATION & CERTIFICATIONS

**MSc Data Science, The University of Sheffield**
September 2022 – January 2024 | Grade – **Distinction**

**Bachelor's in Computer Science and Engineering, CUSAT**
August 2016 - July 2020 | Grade - **First Class**

**Azure Solutions Architect Expert Certification**
February 2024

**Power BI Data Analyst Associate**
September 2021

**IBM Data Science Professional, Coursera**
July 2020

## RESEARCH PROJECTS

**Adversarial Exploits: Classification of AI-generated human faces (Dissertation)**

- Formulated a novel architecture combining CNNs and Transformer networks for robust **classification** of **AI-generated faces**, achieving **83% accuracy** on challenging datasets.
- Constructed advanced **data augmentation techniques**, including **style transfer** and progressive growing of GANs, to enhance model generalization and boosting validation accuracy by 12% on limited training datasets.
- Orchestrated a self-supervised pretraining approach using contrastive learning, reducing the need for large, labeled datasets and improving model performance by 15%.

**RAG-based Chatbot for Sentiment aware Document Retrieval**

- Designed and launched a **Retrieval-Augmented Generation (RAG)** chatbot leveraging **LangChain** for conversational orchestration and **ChromaDB** for semantic vector-based document retrieval.
- Integrated **sentiment analysis** using transformer-based models to adjust response tone and content based on user emotions, increasing user engagement metrics by 18% and reducing negative feedback by 22%.
- Leveraged a modular pipeline for ingesting and embedding **structured** and **unstructured data**, allowing real-time querying over documents and reports.
- Mitigated **LLM hallucinations** using LangChain's Refine and Map-Reduce strategies for grounded answer generation, improving factual consistency and reducing off-context responses.

**Predictive Modeling of Building Energy Demand Using XGBoost & LightGBM**

- Devised an ML model to predict **building-level energy consumption** using historical **meter readings**, **weather patterns**, and **building characteristics** from a public dataset of over 1,400 buildings.
- Incorporated time-series features like **lag metrics**, **moving averages**, and **seasonal encodings** to enhance model performance by 12%.
- Fine-tuned ensemble models (**XGBoost**, **LightGBM**, **CatBoost**), reducing RMSE by up to 18% across multiple meter types (electricity, chilled water, steam, hot water).
- Conducted exploratory data analysis (EDA) to uncover **consumption trends** and **peak usage periods** by region and building type.
- Created interactive dashboards using **Plotly** and **Tableau** to visualize forecasts.

**F1nalyse – Formula 1 Analytics & Machine Learning Platform**

- Developed an advanced Formula 1 analytics platform that combines real-time race data with machine learning models for performance prediction and strategy optimization
- Implemented ML models for race outcome prediction, analyzing historical data patterns and real-time race conditions to forecast race results and driver performance
- Built interactive visualizations using Recharts and custom ML model outputs to display predictive analytics and performance trends
- Developed a modular architecture that separates ML model training, inference, and frontend visualization components
- Implemented data preprocessing and feature engineering pipelines for transforming raw F1 data into ML-ready datasets

## SKILLS

Programming: **Python, R, SQL**
Machine Learning & Deep Learning: **TensorFlow, Keras, PyTorch, Scikit-learn, XGBoost, LightGBM**
Big Data: **Apache Spark, Hadoop, Hive, PySpark, Databricks**
Data Engineering & Pipelines: **Airflow, Kafka, Snowflake, dbt, ETL & Feature Engineering**
DevOps & MLOps: **Docker, CI/CD (GitHub Actions), MLflow, Kubernetes, AWS SageMaker**
Visualization: **Tableau, Power BI, Plotly, Dash, Streamlit**
Cloud: **AWS, Azure, GCP**