

Soumyadip Ghorai

Data Scientist | Bangalore, India

+91 7872817970 | ✉ ghorai.soumyadip33@gmail.com | [in LinkedIn](#) | [GitHub](#) | [Leetcode](#) | [Portfolio](#)

Summary: MSc + BS Data Science graduate with 3+ years of experience building end-to-end AI/ML solutions driving business impact across telecom, pharma, manufacturing, consulting, and sports. Expertise in LLM, RAG, ML, data pipelines, stakeholder management, and cross-functional collaboration. 3X Kaggle Expert and mentor to 200+ students, passionate about building scalable AI solutions.

Work Experience

GE Vernova

Data Scientist — Bangalore

JUN 2025 - PRESENT

- Built **HyDE**-based advanced RAG tool, cutting plant application onboarding from *6+ weeks* to *under 2 weeks* for new configurators.
- Developed *LangGraph* based agent with **LLM+SLM architecture**, enabling talk-to-data across multiple real-time data sources.
- Achieved *97% text classification* accuracy using **SGD + Logistic** and *LoRA-finetuned Llama3.2:1B*, reducing LLM hallucinations.
- Created end-to-end real-time **Kafka to Druid** pipeline, scaling to *300+* control loops, powering dashboards tracking *40+ live KPIs*.

Tech: Python, SQL, Machine learning, LLM, SLM, RAG, LangGraph, Apache Druid, Kafka, Git, Streamlit, Docker, Vector Database

KPMG

Data Scientist — Bangalore

JAN 2024 – JUN 2025

- Enhanced algorithm to parse and chunk PDFs, implemented **RAG** to deliver coherent responses, and cut processing time by *40%*.
- Designed and developed an automated tool for time series forecasting using **GPT 3.5** and **Prophet** automating *80%* of the work.
- Integrated open-source LLMs like **Llama-3** seamlessly into the existing architecture resulting in *30% reductions* in operating costs.
- Optimized inventory parameters using Monte Carlo Simulation, achieving target service level, and reducing cost by *\$500k - \$750k*.

Data Scientist Trainee — Bangalore

FEB 2023 – MAY 2023

- Completed training in Python, SQL, ML, FastAPI, GenAI, and Power Apps, and achieved *92%* accuracy in KPMG's ML hackathon.
- Created Power Apps prototype for leadership demonstrating end-to-end data pipeline automating *5+* dashboard workflows.

Tech: Python, SQL, Machine learning, Large Language Models, RAG, Optimization, Fast API, Azure ML, Bitbucket, Git, Chroma

Tweek Labs

Data Science Intern — Bangalore

MAR 2022 – AUG 2022

- Formulated aggregated scoring algorithm to rank *500+ players* according to their stats to help organizations create leaderboards.
- Streamlined data storage pipeline via Google Sheets API, and built an interactive dashboard using **Meta Base** to track *15+ KPIs*.
- Applied **KNN** and **stacked model** to predict ground contact with an average accuracy of *11ms* to get leg parameters accurately.

Tech: Python, Meta Base, Google Sheets, Machine Learning, Google API, PyCharm, GitHub

Ericsson

Summer Intern — Kolkata

SEP 2021 – OCT 2021

- Root Cause Analysis:** Predict the root cause and recommend possible resolutions to reduce the overall processing time by *50%*.
- Wrote a generalized parser to parse XML log files, and create clean JSON files to upload on elastic search using **Python** and **bs4**.

Tech: Python, VS Code, Beautiful Soup, MS Office

Personal Projects

Milestone.ai — [\[GitHub\]](#) [\[App Link\]](#)

OCT 2024 – DEC 2024

- AI-powered project tracker with **RBAC**, enabling instructors to track student progress and analyse project code and project reports.
- Utilized **LLM** to generate structured milestones, checklists, descriptions, and deadlines for each project from project details pdf.
- Added features to fetch code using **GitHub API** with *50% coverage*, used AI to analyse code, and generate *personalized feedback*.

Tech: Python, Fast API, Vue, Llama-3, SQLite, HTML, CSS, Bootstrap, LangChain JS, RBAC, Figma, JIRA, git, GitHub API

Education

- BS in Data Science and Applications**
- MSc Data Science**
- BSc Statistics**

Indian Institute of Technology, Madras

Christ University, Bangalore

University of Calcutta

2021 – 2025

2021 – 2023

2018 – 2021

Skills

Languages/frameworks: Python, SQL, Tensorflow, PyTorch, scikit-learn, Fast API, Flask, LangChain, LangGraph, Streamlit

Machine Learning Skills: Regression, Classification, Clustering, EDA, Natural Language Processing, Transformers, ANN, RNN, LoRA

Tools/technologies: Azure ML, Git, Apache Druid, Tableau, GCP, AB testing, Kafka, Docker, Kubernetes, Ollama, Chroma

Achievements and PORs

- 3X Kaggle Expert**, featured Streamlit dev, **975+** LeetCode problems, led ML team at **GDSC**, and **2X district athletic champion**.
- Solo winner of 2 consecutive *Hour-of-Code* at **KPMG India**, got featured at **GE Vernova Users Conference 2026** for AI prototypes.