Sayan Banerjee

sayan112207@gmail.com | 9475979608 | sayan112207.github.io | linkedin.com/in/sayan18 | github.com/sayan112207

Summary

Analytical professional focused on evaluating and improving business outcomes through rigorous quantitative analysis. My experience spans the full data pipeline - from data collection and cleaning to building statistical algorithms that drive decision-making. I thrive in environments requiring detailed documentation and problem-solving, ensuring that large internal and external data sets are not only accurate but are communicated in a concise way to support key business goals.

Experience

Data Analyst, WorkIndia, Bangalore — May 2025 – Present

- Increased **Daily Active Users (DAU) by 17%** through behavioral cohort segmentation and targeted outreach strategy.
- Designed and executed controlled A/B experiments on 1M daily website traffic with equal test-control allocation over 14 days, achieving an **8% lift in Clicks-to-Leads at 99% statistical confidence**.
- Led continuous A/B tests across Email and WhatsApp channels, boosting downloads by 34% and 26% YoY.
- Reduced **Cost-per-DAU from 14.5 to 9.7 (33% reduction)** and saved **$1,200 in server costs** by designing time-affinity notification strategies using historical interaction data.
- Wrote complex SQL queries (joins, aggregations, window functions) to analyze retention trends, recency-frequency metrics, and campaign performance across segmented user cohorts.
- Built and maintained cron-driven ELT workflows using SQL and PySpark to automate segmentation pipelines and KPI tracking dashboards.
- Ensured reporting accuracy and data integrity in production environments, proactively mitigating risks from ETL and workflow failures.

Technical Research Analyst, Hevo Data, Bangalore — Jun 2024 – Dec 2024

- Investigated and resolved issues across **80+ distributed ETL pipelines**, improving data reliability for downstream analytics and business reporting.
- Built a Snowflake-based pricing tool used by sales and marketing teams, contributing to a **9% increase in qualified leads**.
- Performed log-level and metric-driven root cause analysis to support system stability and safeguard reporting accuracy.

Research and Development Intern, Samsung Remote — Nov 2023 – May 2024

- Processed and analyzed **32,000+ documents** using scalable data pipelines for structured information extraction.
- Developed transformer-based summarization models, balancing model accuracy, throughput, and deployment efficiency

Education

KIIT University, B.Tech in Computer Science — Sep 2021 – Apr 2025

- **CGPA**: 8.9/10.0
- **Coursework:** Machine Learning, Database Management Systems, Distributed Systems, Cloud Computing

Skills

- **Programming & Analytics:** Python, SQL, PySpark, Pandas, NumPy
- **Data Analysis:** A/B Testing, Cohort Analysis, KPI Tracking, Statistical Validation, EDA, Hypothesis Testing, Experiment Design
- **Databases:** MySQL, PostgreSQL, Snowflake, MongoDB, Athena
- **Tools & Infrastructure:** Git, Linux, JIRA, Confluence, Kubernetes, Jenkins

Academic Projects

Perplexa, KIIT — Jan 2025 – April 2025

- Integrated Google OAuth with MongoDB-backed authentication and Captcha verification, ensuring **99.9% uptime**.
- Built a FAISS-based RAG with real-time web context, **reducing hallucination by 7%** and maintaining **<150ms latency**.
- Deployed via SSL-secured GitHub webhooks to Streamlit Cloud, scaling to **500 concurrent queries/min**.

Vehicle Movement Analysis and Insight Generation, Intel Unnati — May 2024 – Jul 2024

- Leveraged YOLOv8 for real-time vehicle detection and tracking, achieving **<100ms inference latency** and **95% accuracy**.
- Used PaddleOCR for license plate recognition, processing 60+ frames/min with **92% accuracy** under variable conditions.
- Optimized edge inference using OpenVINO, realizing a **3x speed boost** for efficient traffic monitoring.

Text-to-SQL, KIIT — Dec 2023 – Feb 2024

- Fine-tuned Star-Coder2-3b on SQL data using PEFT LoRA, achieving a **28% improvement** in query-to-code accuracy.
- Leveraged vectorization and token embedding to cut **input processing time by 35%**, in a resource-constrained environment.
- Applied 4-bit quantization using bitsandbytes, **lowering memory usage by over 20%**.