SUNNY KUMAR

Data Analyst & AI/ML Data Associate | MCA Final Year (2026 Batch)

sunnykr155415@gmail.com | +91 8864091706 |

LinkedIn: linkedin.com/in/sunny-kumar-01208a268 | GitHub: github.com/Sunnykumar1554 | Portfolio: sunnykumar-portfolio.netlify.app

PROFESSIONAL SUMMARY

Detail-oriented Data Analyst and AI enthusiast pursuing MCA (Final Year, 2026 Batch) with hands-on experience in data labeling, annotation quality control, Python, SQL, Power BI, and NLP-based AI systems. Proven ability to handle multi-modal data (text, speech, image, audio, video), follow complex labeling guidelines, maintain strict confidentiality and customer privacy, deliver high-quality labeled datasets, and apply structured problem-solving under ambiguous conditions. Fluent in Hindi and English; strong business writing and reading comprehension skills. Eager to support Amazon's responsible AI/ML development as an ML Data Associate.

SKILLS & COMPETENCIES

Programming & Data: Python, Pandas, NumPy, SQL, MySQL, PostgreSQL, Statistics, Prophet, LSTM, XGBoost, Optuna

AI / ML & NLP: Large Language Models (LLMs), Data Labeling & Annotation, Prompt Evaluation, Multi-modal Data (Text, Speech, Image, Audio, Video)

Tools & Platforms: Power BI, Tableau, Excel, Google Colab, VS Code, ETL Pipelines, Streamlit, MLflow, Evidently AI, Docker, Kubernetes, GitHub Actions, Prometheus, Grafana, AI Tools

Soft Skills: Attention to Detail, Analytical Thinking, Structured Problem-Solving, Judgment Under Ambiguity, Hindi & English Fluency, Teamwork, Adaptability

INTERNSHIP EXPERIENCE

Zidio Development — **Data Science & Analysis Intern** Bangalore, India | 2026 (2 Months)

- Performed exploratory data analysis (EDA) on real-world datasets using Python (Pandas, NumPy, Matplotlib, Seaborn), identifying key patterns and delivering actionable business insights.
- Cleaned, labeled, and structured multi-format datasets (tabular, text, speech) following strict labeling guidelines — maintaining confidentiality of sensitive data in line with customer privacy standards.
- Built interactive Power BI and Tableau dashboards to visualize KPIs and operational metrics for stakeholder reporting.
- Documented data processing workflows, proposed process improvements, and maintained high accuracy standards under deadline-driven conditions.
- Collaborated with cross-functional team members, demonstrating adaptability across multiple task types simultaneously.

PROJECTS

RetailPulse — AI-Powered Customer Analytics & Demand Forecasting Platform | Python, Prophet, LSTM, XGBoost, Optuna, Streamlit, MLflow, Docker, Kubernetes | https://retailpulse1554.netlify.app/

- Built an end-to-end MLOps platform processing 1M+ retail transactions, using a Prophet + LSTM hybrid ensemble to cut demand-forecast error (MAPE) to 24.1%.
- Developed RFM-based customer segmentation (K-Means, DBSCAN) and an XGBoost churn classifier tuned with Optuna, achieving 0.94 ROC AUC.
- Shipped a 5-page Streamlit dashboard with MLflow tracking, Evidently drift detection, Docker/Kubernetes deployment, GitHub Actions CI/CD, and Prometheus/Grafana monitoring.

MediAssist — AI Healthcare Data & Annotation Platform | Python, NLP, LLMs, Llama API, Data Quality

- Designed and developed an AI-powered healthcare assistant using NLP and Large Language Models (LLMs) to process and annotate symptom-based medical text data across multiple modalities.
- Curated and labeled high-quality human insight data from medical text inputs, implementing safety checks, disclaimers, and escalation logic — mirroring Amazon's responsible AI data pipeline requirements.
- Integrated Llama API for generative response handling; applied rigorous quality control protocols to ensure labeled output accuracy met defined KPIs.
- Identified and resolved annotation inconsistencies by analyzing root causes and proposing improvements to the Standard Operating Procedure (SOP).

Hospital ER Data Analytics — Full-Stack Solution | Python, MySQL, Power BI, Pandas

- Automated data cleaning and feature engineering on 500+ patient records using Python Pandas; applied labeling and classification logic to structured healthcare data.
- Wrote advanced MySQL queries for readmission rate tracking and revenue analysis, demonstrating proficiency in complex data document interpretation.
- Built an interactive Power BI dashboard visualizing patient flow, diagnosis trends, and department KPIs — showcasing data visualization across structured datasets.

Real-Time Kafka Streaming Pipeline | Apache Kafka, Python, ETL

- Built a distributed Kafka producer-consumer ETL pipeline processing 10,000+ messages/sec; designed for high-throughput, low-latency multi-format data handling.

Big Data Analysis — Global E-Commerce Transactions | Apache Spark, PySpark, Data Visualization

- Analyzed 1M+ transaction records using Apache Spark; reduced processing time by 60% with distributed PySpark aggregation and delivered visualized sales and revenue insights.

EDUCATION

Tula's Institute — **Master of Computer Applications (MCA) — Final Year** Dehradun, Uttarakhand | Expected 2026

L.N. Mishra Institute of Economic Development & Social Change — **Bachelor of Computer Applications (BCA)** Patna, Bihar | 2024 | CGPA: 7.42

CERTIFICATIONS

- Google Data Analytics Professional Certificate — Coursera (Google)
- Microsoft Power BI Data Analyst (PL-300) — Microsoft
- Web Development — SimpliLearn / SkillUP