Ansh Kumar Dev

520-599-5589 | anshkumardev@gmail.com | linkedin.com/in/ansh-kumar-dev/ | github.com/Anshkumardev

EDUCATION

University of Arizona — Tucson, AZ
Master of Science in Data Science — *Aug. 2023 – May 2025*

Gautam Buddha University — Greater Noida, India
B.Tech. in Artificial Intelligence — *Aug. 2019 – May 2023*

EXPERIENCE

AI Engineer Intern — Jun 2025 – Present
Right Skale — *Pleasanton, CA*
- Designed and automated pipelines for ingesting and validating **SEC 10-Q and balance-sheet data**, ensuring consistency and alignment with **regulatory reporting** standards.
- Monitored and documented **data quality issues** across ingestion streams; collaborated with product owners to escalate critical risks and define **remediation plans**.
- Worked with business and engineering teams to define **Key Data Elements (KDEs)** and ensure traceability across system components via metadata tagging and data lineage tracking.
- Applied **data governance controls** to containerized workflows using GitHub Actions, IAM, CloudWatch, and X-Ray to support **SOX-compliant** deployments.
- Implemented automation to reduce manual QA in pipeline validation, driving over **90% automation** in end-to-end ingestion with issue logging for anomalies.

Data Scientist (Graduate Research Assistant) — Aug 2024 – Jan 2025
University of Arizona — *Tucson, AZ*
- Developed an ETL pipeline in **Python (Pandas, SQL)** to fetch and clean structured and unstructured sensor data; enforced **data validation** and transformation rules to prevent pipeline inconsistencies.
- Led **gap assessments** and root cause analysis on sensor feeds with missing attributes; implemented compensating controls for handling incomplete records.
- Generated metadata documentation for model features and mapped **data lineage** across acquisition, feature engineering, and modeling layers.
- Collaborated with academic advisors and domain SMEs to present issue summaries, validations, and corrective actions for experimental forecasting.

AI Engineer Intern — Apr 2022 – Jul 2022
Merkletree Technologies — *Delhi, India*
- Tracked model performance failures and escalated **data quality issues** within biometric pipelines, implementing logging strategies to support reproducibility.
- Deployed models with **Docker Kubernetes** using pre-validation scripts to check field-level integrity and **enforce input constraints**.
- Defined compensating controls for production anomalies; partnered with backend and mobile teams to document corrective measures and align release standards.
- Maintained Git-based lineage of model versioning and data drift indicators to support traceability in downstream releases.

PROJECTS

Healthcare Accessibility & Risk Mapping Platform | *Project Link* — Apr 2025
- Built a distributed **GeoPandas + OSMnx** pipeline unifying Geospatial, socioeconomic & clinical data across 3,000+ U.S. census tracts to identify underserved areas for policy action.
- Developed a **LangChain** multi-agent LLM workflow with **OpenAI API** and **FAISS** RAG and a **Streamlit** choropleth dashboard, enabling real-time Q&A and evidence-based insights.

E-commerce Analytics Lakehouse | *Project Link* — Mar 2025
- Engineered an **Azure Data Factory to Databricks** (**PySpark**) ETL ingesting 1.2 TB/day of order & clickstream data to power near-real-time KPI dashboards.
- Modeled star-schema views in **Synapse SQL** and built live **Power BI** reports, Reduce manual reporting by 90%

TECHNICAL SKILLS

Languages/ML: Python (Pandas, NumPy, Scikit-learn, TensorFlow, PyTorch, OpenCV), R, SQL, C++
Models: RoBERTa, SVM, Decision Trees, K-Means, PCA, Forecasting & Statistical Analysis
Cloud & MLOps: AWS (EKS, EC2, S3), Azure (ADF, Synapse, Databricks), GCP, Docker, Kubernetes, Spark, CI/CD, MLflow
Tools & APIs: Flask, REST APIs, TensorBoard, Git, Linux, Power BI, Streamlit