Debankur Saha

Senior Machine Learning Engineer

MLOps · ML Platform Engineering · Kubernetes · Cloud & On-Premise Infrastructure

Hyderabad, India | debankur_123@yahoo.com | 8981532992 | github.com/DebankurS | linkedin.com/in/debankur-saha | debankurs.com

PROFILE

Machine Learning Engineer with 11 years of experience architecting and operating production ML platforms across cloud and on-premise infrastructure, with experience spanning fintech, banking, and enterprise SaaS. Specialized in ML system design, CI/CD and deployment automation, Kubernetes-based ML platforms, and model lifecycle management (training, inference, monitoring, drift detection) under real-world latency, cost, and reliability constraints. Builds the infrastructure foundations — including GenAI/LLM serving — that let ML teams ship and operate models reliably at scale.

WORK EXPERIENCE

Senior Software Engineer II - Machine Learning
Feb 2023 – Present

Bottomline Technologies

- Own the MLOps process for the Data Science team, defining standard ML system design and patterns (training, inference, monitoring, CI/CD) that reduce deployment time and operational incidents.
- Lead and mentor a 3-person team building an ML-based vendor deduplication and search system for Bottomline's flagship B2B payments network (600,000+ verified businesses), deploying models via Snowflake Snowpark — improving match accuracy, reducing duplicate vendor entries, and cutting turnaround-approval times from weeks to hours.
- Design and deploy ML workloads onto org-wide multi-tenant Kubernetes clusters spanning multiple regional data centers, using namespace isolation to balance security, latency, and cost constraints.
- Modernized org-wide Python CI/CD pipeline on GitLab CI to support uv and pyproject.toml while maintaining full backward compatibility with legacy setup.cfg-based repositories across the organization.
- Built a provider-agnostic AI Gateway (AWS Bedrock, OpenAI-compatible APIs) with dual auth support (Keycloak/JWT and static API keys), rate-limiting, and observability, serving as the platform layer for the org's GenAI initiatives.

Machine Learning Engineer
Sep 2021 – Jan 2023

TDCX

- Owned ML components for the Digital Lab on Google Cloud Platform (GCP), leveraging Vertex AI while identifying architectural gaps and introducing MLOps best practices across projects.
- Designed containerized inference architectures for ML services using Docker and REST APIs, standardizing deployment across teams.
- Implemented DevOps best practices in non-ML projects, optimizing server usage and reducing infrastructure costs.
- Led cost-optimization initiatives by replacing paid AI services with open-source alternatives, cutting recurring vendor spend.
- Built in-house translation and facial feature recognition services using open-source models and libraries.

Data Scientist / Computer Vision Engineer
Jun 2020 – Sep 2021

Axis Bank

- Designed an automated PAN validation system combining OCR, NLP validation logic, and human-in-the-loop fallback, achieving 85–90% accuracy while reducing manual review load.
- Automated image-to-video validation for KYC verification during customer onboarding.
- Built verbal KYC authentication solution using Azure Speech to Text with regional language support.
- Co-created a dispute resolution solution for FASTag operations, using Azure Cognitive Services' Computer Vision OCR to extract and parse toll receipts and validate duplicate payments.

Developer / Data Scientist
Sep 2015 – May 2020

Tata Consultancy Services

- Built a gesture-based rating system using transfer learning on MobileNetV2, fine-tuned for custom gesture recognition and deployed as Flask APIs on Azure App Service.
- Built a TensorFlow-based email classifier for service ticket triage; introduced Dockerization to eliminate deployment dependency conflicts.

SKILLS

MLOps & Model Lifecycle: Model lifecycle design, Model deployment, Training & inference pipelines, Model monitoring & drift detection, Feature engineering & evaluation, MLFlow, Vertex AI, SageMaker

Platform & Orchestration: Kubernetes, Docker, Helm, ArgoCD, CI/CD (GitLab), Hashicorp Vault

Cloud & On-Premise Infrastructure: AWS, Google Cloud Platform (GCP), Microsoft Azure, Self-hosted / on-premise data centers

Data Platforms & Compute: Snowflake, Snowpark, CUDA

ML & AI: Deep Learning (CV, GenAI, LLMs), TensorFlow, PyTorch, Scikit-Learn, Huggingface Transformers, LLM inference (vLLM, SGLang, TensorRT-LLM)

AI-Assisted Development: Claude Code, GitHub Copilot

Programming Languages: Python, JavaScript (Node.js)

EDUCATION

Netaji Subhash Engineering College — B.Tech, Electronics and Communication Engineering, 2011–2015, Kolkata, India
Calcutta Airport English High School — West Bengal Council of Higher Secondary Education, 2009–2011, Kolkata, India

CERTIFICATES

- AWS Certified AI Practitioner — Amazon Web Services / AWS
- Machine Learning Basic Nanodegree — Udacity
- Deep Learning Nanodegree — Udacity
- Computer Vision Nanodegree — Udacity
- Deep Reinforcement Learning for Enterprise — Udacity
- Machine Learning in the Enterprise — Coursera
- Production Machine Learning Systems — Coursera

AWARDS

- MVP Award (Data and ML Team), Bottomline Technologies — 2025
- Udacity PyTorch Scholarship from Facebook AI — 2018–2019
- Udacity Secure and Private AI Scholarship from Facebook AI — 2019

OPEN-SOURCE CONTRIBUTION

PyTorch Contributor

- Restored functionality of the official PyTorch CUDA 12.8 devcontainer by hotfixing the cuda-keyring dependency architecture and standardizing apt package naming conventions.
- Resolved critical repository linting blocks by debugging and patching internal validation rules within the project's build infrastructure (.spin/cmds.py).
- Collaborated with core maintainers to navigate CI/CD pipeline irregularities and successfully merge infrastructure fixes.