Machine Learning Engineer

AI / ML | SMB | Enterprise

We are looking for a Machine Learning Engineer to bridge the gap between data science research and production engineering at [Company Name]. You will take ML models from prototype to production, build robust training and serving infrastructure, and ensure models perform reliably at scale. This role requires strong software engineering fundamentals combined with deep understanding of ML systems and their unique operational challenges.

Key Responsibilities

- Design and implement end-to-end ML pipelines including data preprocessing, feature engineering, model training, evaluation, and deployment
- Build and maintain model serving infrastructure that handles real-time and batch prediction workloads with low latency and high reliability
- Develop feature stores and feature engineering pipelines to provide consistent, reusable features across training and serving environments
- Implement model monitoring, drift detection, and automated retraining pipelines to keep models performant over time
- Optimize model performance for production constraints including latency, throughput, memory, and cost
- Collaborate with data scientists to productionize research models, translating Jupyter notebook prototypes into maintainable, tested code
- Build and maintain MLOps tooling and infrastructure including experiment tracking, model registry, and CI/CD for ML

Required Skills & Experience

- 4+ years of software engineering experience with at least 2 years focused on ML systems in production
- Strong Python programming skills with emphasis on production-quality code (testing, error handling, logging, documentation)
- Experience with ML frameworks: PyTorch, TensorFlow, or JAX for model development and training
- Hands-on experience deploying models to production using serving tools like TorchServe, TensorFlow Serving, Triton, or custom APIs
- Familiarity with MLOps tools: MLflow, Weights & Biases, Kubeflow, or similar experiment tracking and pipeline orchestration
- Experience with cloud ML services (AWS SageMaker, GCP Vertex AI, or Azure ML)
- Strong understanding of Docker, Kubernetes, and microservice architectures for model deployment
- Knowledge of data processing at scale using Spark, Dask, or Ray

Nice-to-Have

- Experience with LLM fine-tuning, prompt engineering, or RAG (Retrieval-Augmented Generation) systems

- Familiarity with model optimization techniques: quantization, pruning, distillation, ONNX conversion
- Experience with feature stores (Feast, Tecton) and real-time feature serving
- Knowledge of A/B testing infrastructure for model evaluation in production
- Experience with GPU cluster management and distributed training

Tech Stack

Python PyTorch TensorFlow Docker Kubernetes MLflow AWS SageMaker Spark Airflow FastAPI Redis Triton Inference Server

What We Offer

- Competitive salary and equity at [Company Name]
- Access to dedicated GPU clusters and cutting-edge ML infrastructure
- Budget for ML conferences (NeurIPS, ICML, MLSys) and continuous learning
- Comprehensive health, dental, and vision insurance
- Flexible remote work with quarterly in-person team gatherings
- Opportunity to work on ML systems that serve millions of users at scale

Interview Process

1. Recruiter phone screen (30 min) — background, ML experience depth, and logistics
2. Technical phone screen (60 min) — coding exercise in Python, ML system design discussion
3. ML system design interview (60 min) — design an end-to-end ML pipeline for a realistic scenario, discuss trade-offs in serving, monitoring, and retraining
4. Coding deep-dive (60 min) — implement a data processing or model serving component with emphasis on production quality
5. Hiring manager and team conversation (60 min) — collaboration style, past project deep-dive, and career goals
6. Reference checks and offer