# Param Madan

Boston, MA, USA | +1 617-259-8214 | madanparam18@gmail.com | LinkedIn | GitHub

## Summary

**AI/ML Engineer** with **3+ years of experience** designing, building, and deploying scalable **machine learning systems** and intelligent applications across high-growth tech, HR-tech, and enterprise environments. Hands-on expertise with **Python**, **Apache Spark**, **PyTorch**, **TensorFlow**, **FastAPI**, **AWS**, and **Kubernetes**, with a strong foundation in **Generative AI**, **Large Language Models (LLMs)**, **Retrieval-Augmented Generation (RAG)**, **MLOps**, **Apache Kafka**, **MLflow**, and **Databricks**. Proven track record delivering production-ready AI solutions that process millions of daily transactions, reduce operational costs, and accelerate decision-making for cross-functional teams. Experienced in **end-to-end ML life-cycle management** from **feature engineering** and **model training** to **CI/CD pipelines**, **experiment tracking**, and **real-time model monitoring** at scale. Adept at collaborating with engineering, product, and data science stakeholders to translate complex business requirements into cloud-native, enterprise-grade **AI/ML platforms**.

## Experience

**AI/ML Engineer** Mar 2026 – Present
*Uber* USA

- Architected **real-time demand forecasting pipelines** using **Python**, **PySpark**, and **Apache Spark** across 100+ metropolitan zones, enabling accurate driver supply planning during peak demand.
- Engineered pricing models using **Apache Kafka**, **XGBoost**, and **Feature Stores**, deploying **12+ production models** for fare estimates and surge recommendations on high-volume transactions.
- Optimized **online inference services** using **PyTorch**, **Docker**, and **Kubernetes** to process **800K+ prediction requests daily** with **sub-second response times** for rider and driver applications.
- Orchestrated ML workflows using **Apache Airflow**, **MLflow**, and **Amazon SageMaker** with 6 cross-functional stakeholders to streamline model validation and production release cycles.
- Championed generative AI initiatives leveraging **LangChain**, **Vector Databases**, and **LLM frameworks** with 2 engineering teams to deliver support assistants and knowledge retrieval workflows.

**AI/ML Consultant** Jul 2024 – Dec 2024
*Massachusetts Institute of Technology - Spinout (AutonomUS)* USA

- Architected a battlefield surgical robotics system to automate wound detection across 14 trauma scenarios using **real-time ultrasound imaging** and **CNNs** via cross-functional clinician workshops.
- Optimized deep learning **U-Net models** within **PyTorch** and **CUDA** to eliminate 3 probabilistic risk conditions by anchoring code changes in thorough peer reviews.
- Orchestrated production pipelines tracking **model drift** across **25,000 validation ultrasound frames** by translating technical risks directly into concrete milestones for executive project stakeholders.
- Overhauled the end-to-end architecture planning and deployment of AI models to hit a **0.8-second processing latency threshold** by unifying engineering teams on workflows.
- Championed system deployment to satisfy 4 military medical standards by building **compliance verification frameworks** rooted in technical ingenuity and cross-team peer alignment.

**Software Developer** Jun 2021 – Jul 2023
*Dell Technologies* India

- Developed backend modules using **Python**, **FastAPI**, and **PostgreSQL** managing **25K+ enterprise devices** across 8 operational workflows for internal asset management applications.
- Engineered data processing components using **Pandas** and **SQL** to collect, clean, and transform **500K+ device events** for reporting dashboards and predictive analytics.
- Integrated prediction services with **Scikit-learn** and **XGBoost** into **backend APIs** enabling hardware health assessments across 5 infrastructure monitoring systems.
- Automated build and deployment using **Docker** and **Jenkins** for **15+ Python services** across development, testing, and production environments.
- Implemented ML integration workflows by validating and deploying **6+ predictive models** into enterprise support platforms, coordinating API contracts with data science teams.

## Projects

**LLM-Powered Multi-Agent RAG System** | *LangChain, Pinecone, FastAPI, AWS ECS* 2024

- Built a **multi-agent RAG pipeline** with **dynamic document retrieval** over **50K+ knowledge base entries**, achieving **87% response accuracy** using **semantic chunking** and **reranking strategies**.
- Deployed as a containerized **FastAPI microservice** on **AWS ECS** with **auto-scaling**, serving **5K+ queries/day** with under **200ms latency**.

**Real-Time Fraud Detection Platform** | *Kafka, Spark Streaming, XGBoost, Kubernetes* 2023

- Engineered a **streaming fraud detection system** processing **10K+ transactions/second** using **Apache Kafka** and **Spark Streaming** with **XGBoost ensemble models** achieving **96% precision**.
- Deployed on **Kubernetes** with **MLflow experiment tracking**, reducing **false positive rate by 23%** over baseline models through iterative **feature engineering**.

**End-to-End MLOps Platform** | *MLflow, Airflow, Docker, AWS SageMaker, Terraform* 2024

- Designed a fully automated **MLOps platform** orchestrating **model training, versioning, and deployment** using **MLflow** and **Apache Airflow**, cutting **release cycle time** from **3 days** to under **4 hours**.
- Provisioned scalable **cloud infrastructure** on **AWS SageMaker** with **Terraform**, enabling **one-click model promotion** across dev, staging, and production environments with **drift monitoring alerts**.

**Intelligent HR Chatbot with Semantic Search** | *Hugging Face, FAISS, FastAPI, React.js* 2024

- Built a domain-specific **conversational AI assistant** using **Hugging Face sentence-transformers** and **FAISS vector indexing** over **200K+ HR policy documents**, reducing **ticket volume by 35%**.
- Exposed inference via a **FastAPI REST layer** with **JWT authentication**, integrated into a **React.js portal** used by **8K+ employees** across **4 business units**.

**Distributed Time-Series Forecasting Engine** | *PySpark, Prophet, Databricks, Delta Lake* 2023

- Developed a **distributed forecasting engine** on **Databricks** using **PySpark** and **Prophet** to generate **30-day demand forecasts** across **500+ SKUs**, improving **inventory accuracy by 19%**.
- Stored **feature snapshots** and model outputs in **Delta Lake** with **time-travel support**, enabling **reproducible backtests** and audit-ready **compliance reporting**.

## Technical Skills

- **Programming Languages:** Python, Java, R, SQL, TypeScript, JavaScript, Go, Bash
- **AI & Machine Learning:** Scikit-learn, TensorFlow, PyTorch, XGBoost, Pandas, NumPy, Hugging Face, NLP, Feature Engineering, Predictive Modeling, Deep Learning, Model Evaluation
- **Generative AI & LLM Engineering:** LLMs, RAG, LangChain, Prompt Engineering, AI Agents, Semantic Search, Embedding Models, Vector Databases (Pinecone, FAISS, ChromaDB), LLMOps, VLLM
- **Data Engineering:** Apache Spark, PySpark, Apache Kafka, Apache Airflow, Databricks, Hadoop, Dask, ETL/ELT Pipelines, Batch & Real-Time Data Processing, Data Modeling
- **Cloud, DevOps & MLOps:** AWS (EC2, S3, Lambda, ECS, EKS, SageMaker, Glue, Redshift), GCP (BigQuery, Cloud Run), Azure (Azure ML, Synapse, AKS), Docker, Kubernetes, Terraform, MLflow, CI/CD
- **Backend & Software Engineering:** FastAPI, Flask, Django, Spring Boot, Node.js, REST APIs, GraphQL, Microservices, Event-Driven Architecture, OAuth 2.0, OOP, System Design, Agile
- **Databases & Storage:** PostgreSQL, MySQL, MongoDB, Redis, Snowflake, Neo4j, Oracle, Azure Cosmos DB
- **Developer Tools:** Git, GitHub Actions, Jenkins, Linux, Tableau, Power BI, Alteryx

## Education

**Northeastern University** Boston, MA, USA
Master of Science in Engineering Sep 2023 – Dec 2025