Senior AI/ML Engineer

Job Title: Senior AI/ML Engineer
Experience Level: 6+ Years
Employment Type: Full-Time
Location: Gurugram, Sector 33
Shift Timings: 12:00 PM - 9:00 PM IST

About the Role

We are looking for a hands-on **Senior AI/ML Engineer** who can own the full lifecycle of machine learning solutions – from problem definition and data modelling to training, deployment, monitoring, and continuous improvement.

You should be comfortable working with messy real-world data, designing robust data models & features, building and training models, and shipping them to production with proper MLOps practices. You must also be aware of the current AI/ML landscape (LLMs, embeddings, vector search, modern tooling) and know when to use what.

Key Responsibilities

End-to-End Solution Ownership

- Work with product / domain stakeholders to understand business problems and define ML use cases
- Translate requirements into data & model design, success metrics, and clear technical plans
- Own the full pipeline: data ingestion → cleaning → feature engineering → model training → evaluation → deployment → monitoring

Data Modelling & Feature Engineering

- Design and maintain data models / schemas optimized for analytics and ML training (batch & real-time)
- Perform exploratory data analysis (EDA) and feature engineering to improve signal quality and model performance
- Work closely with data engineering to ensure reliable, well-documented datasets

Model Training & Evaluation

- Build, train, and tune models for tasks such as: prediction, classification, ranking, recommendations, anomaly detection, NLP, etc.
- Use appropriate techniques (traditional ML, deep learning, embeddings, LLMs) based on the problem
- Define and track offline and online metrics; run A/B tests or controlled experiments where applicable

MLOps & Productionization

- Build reproducible training pipelines (e.g., using MLflow, Airflow, Kubeflow, or similar tools)
- Package and deploy models as APIs / microservices or batch jobs, using containers and cloud services
- Implement monitoring, alerting, and logging for model performance, data drift, and system health
- Manage model versions, rollouts, and rollback strategies

AI/ML Architecture & Best Practices

- Evaluate and integrate modern AI tools: vector databases, embedding models, LLM APIs, RAG architectures, etc.
- Ensure solutions follow security, privacy, and compliance best practices (e.g., PII handling, access control)
- Write clear documentation for data flows, models, and services
- Mentor junior engineers / data scientists and contribute to engineering standards and guidelines

Must-Have Skills & Experience

Core Technical Skills (6+ Years)

- **Python Programming:** Strong expertise in ML libraries (pandas, numpy, scikit-learn, PyTorch, TensorFlow)
- **SQL & Databases:** Solid SQL skills and hands-on experience with relational and NoSQL data stores
- **Production ML:** Demonstrated experience shipping end-to-end ML projects to production (not just notebooks / POCs)
- **ML Fundamentals:** Deep understanding of supervised/unsupervised learning, evaluation metrics, overfitting, bias/variance, data leakage, etc.

MLOps & DevOps

- Experiment tracking tools (MLflow, Weights & Biases, etc.)
- Model versioning and packaging (Docker, virtualenv, Conda)
- CI/CD pipelines for ML services
- Infrastructure as Code and containerization best practices

Cloud & Architecture

- Proficiency with at least one major cloud platform:
  - **AWS:** S3, EC2, SageMaker, Lambda, RDS, DynamoDB
  - **GCP:** Cloud Storage, Compute Engine, Vertex AI, Firestore
  - **Azure:** Blob Storage, VMs, Azure ML, Cosmos DB
- API design (REST/GraphQL) and microservice architecture integration
- Understanding of scalability, latency, and cost optimization

Modern AI/ML Landscape Awareness

- Exposure to LLMs & embeddings (OpenAI, HuggingFace, Anthropic, etc.)
- Familiarity with vector search & semantic search platforms (OpenSearch, Elasticsearch, Pinecone, Weaviate, pgvector)
- Ability to make technical trade-offs between classical ML vs deep learning vs LLM-based approaches
- Understanding of cost, latency, and accuracy considerations for each approach

Soft Skills

- **Problem-Solving:** Strong analytical thinking with ability to question requirements and propose better solutions
- **Independence:** Can drive projects from ideation through production deployment with minimal guidance
- **Communication:** Excellent at explaining technical trade-offs and complex concepts to both technical and non-technical stakeholders
- **Collaboration:** Works well with cross-functional teams (product, data engineering, infrastructure, security)

Nice-to-Have Experience

- Data modelling for analytics (star schema, dimensional modelling, data marts)
- Domain expertise in: education, healthcare, finance, e-commerce, or similar industries

- Streaming data platforms (Kafka, Kinesis, Google Pub/Sub) and real-time ML use cases
- Feature store implementation or similar model reuse mechanisms
- Building recommendation systems, anomaly detection pipelines, or NLP solutions
- RAG (Retrieval-Augmented Generation) systems, chatbots, or text summarization
- Experience with model interpretability and explainability (SHAP, LIME, etc.)
- Knowledge of distributed computing frameworks (Spark, Ray, Dask)