Sumeet Khatri

Machine Learning Engineer | Deep Learning| MLOps | System Design

sumeetkhatri1771@gmail.com +919328050693 Pune Linkedin GitHub Portfolio

PROFESSIONAL SUMMARY

Machine Learning Engineer & Software Developer skilled in building end-to-end AI systems and high-performance software solutions. Proficient in Python, C++, MLOps, NLP, and deep learning, with experience in AWS, Azure, and large-scale data engineering. Strong foundation in DSA and System Design, delivering solutions that cut decision-making time by 90% and reduce operational costs by 30%+.

Projects

Building a GPT-Language Model

- Designed and implemented a **transformer-based language model** in **PyTorch**, coding attention mechanisms & tokenization logic from scratch.
- Architected a **training pipeline** that reduced **validation loss by 15%**, generating **context-aware sentences** for downstream NLP tasks.
- Open-sourced the model with documentation, GitHub repo, and deployment on **HuggingFace Spaces** for public access.

Toxic Comment Classification

- Engineered an end-to-end NLP pipeline for toxic comment detection by preprocessing 100k+ text samples using **NLTK, Pandas**, and TF-IDF features.
- Trained & evaluated a **Logistic Regression** model achieving **95% accuracy**, optimizing precision/recall trade-offs for production deployment.

Car Damage Prediction

- Developed a **deep learning image classifier** with **92% accuracy** on **11k+ labeled images**, enabling **real-time damage detection** across multiple datasets.
- Built a modular **data preprocessing-to-model pipeline** in **PyTorch**, improving training efficiency and scalability for production use.

CrewAI Multi-Agent System Prototype

- Engineered a **CrewAI-based multi-agent system** deploying autonomous "crew" agents for collaborative task execution, **improving orchestration efficiency by 35%** and enhancing workflow modularity.
- Implemented configurable agent behaviors and **inter-agent communication protocols**, enabling **dynamic task delegation** and reducing end-to-end task completion time by **~40%** in prototype testing.

Credit Risk Modeling

- Built a **predictive credit risk model** with **94% accuracy**, reducing manual loan reviews by **60%** and decision time from **5 min to 15 sec**.
- Deployed on a **Streamlit + AWS** pipeline for real-time evaluation, enhancing loan officer decision-making.

Predictive Health Insurance Premium Model

- Designed a high-accuracy (>97%) premium estimation model using demographic & medical data, reducing error margin by **10%**.
- Deployed a secure **cloud-hosted app** enabling instant predictions for underwriters.

Professional Experience

Virtual Software Engineer Intern, *-Deloitte Australia*
06/2025 – 07/2025 | India

- Analyzed 10M+ sensor records using PySpark, improving efficiency by 40%.
- Built KPI dashboards with Power BI and automated reports
- Proposed AWS pipeline (Kinesis, Lambda, Redshift) cutting costs by 30%.
- Facilitated workshops and wireframe creation in Figma.

Technical Skills

- **Languages**: Python, C++
- **ML & DL**: Scikit-learn, TensorFlow, NLTK, PyTorch, XGBoost, LightGBM, CNNs, NLP, Transformers
- **MLOps & Deployment**: FastAPI, Streamlit, REST APIs, Git, ML Pipelines, CI/CD Pipelines, Cloud, Model Deployment
- **Data Engineering**: PySpark, Data Preprocessing, Feature Engineering
- **Developer Tools:** Git, GitHub, Docker, VS Code, Google collab
- **Cloud:** Azure, AWS, Google Cloud
- **Foundations**: System Design, object-oriented programming(OOP), Data Structures & Algorithms (DSA)
- **Generative AI:** Feature Engineering, Development AI-Agent, Langchain, Langgraph, Vector Database, LLMs, Retrieval Augmented Generation (RAG), Model Context Protocol (MCP), CrewAI

Education

B.Tech IT, *Indus University*
07/2022 – 07/2026 | Ahmedabad

Certificates/Courses

Deloitte
Virtual internship providing hands-on experience in data processing

C++ DSA
Course focused on DSA concepts using C++

Master in Machine Learning
Completed project-based course covering core ML algorithms, pipelines, and model deployment strategies.

Deep Learning: Beginner to Advanced
Completed a hands-on, project-based deep learning course covering neural networks, CNNs, RNNs, transformers, optimization and deployment techniques using PyTorch for real-world applications.