WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Containerized AI Micro services for Demand Sensing in SMEs: An Internship Experience

Anishi Raj

Department of Electronics and Communication Engineering, Bengaluru 560078,India.

A B S T R A C T:

This paper presents a comprehensive overview of a technical internship focused on the development and deployment of AI-based microservices for small and medium-sized enterprises (SMEs). The internship aimed to transition experimental AI models from Jupyter Notebooks into scalable, production-ready services using containerization, RESTful API frameworks, and CI/CD pipelines. Special emphasis was placed on a demand sensing use case in supply chain management, where advanced machine learning workflows and cloud deployment strategies were implemented. The experience provided practical insights into modern DevOps and MLOps practices, bridging the gap between academic learning and industry-level AI engineering.

Keywords: AI Micro services, Demand Sensing, DevOps, MLOps, FastAPI, Docker, CI/CD, Kubernetes, Supply Chain Forecasting, Cloud Deployment, SMEs

Introduction

During my internship, I worked with a company that helps small and medium-sized businesses adopt AI technologies. My primary role involved converting AI agents—initially built as Jupyter Notebooks—into micro services that could run reliably in production. I was also responsible for ensuring the infrastructure supported continuous deployment and could scale efficiently on cloud platforms.

Company and Department Overview

I worked in the Artificial Intelligence Development department and collaborated with the DevOps, API Integration, and Research teams. The organization followed an agile and collaborative structure, which allowed me to interact across teams. This gave me exposure not only to AI development but also to real-world deployment pipelines and infrastructure planning.

Methodology My tasks focused on transitioning six standalone AI agents into deployable services:

- Microservice Conversion: I used FastAPI to turn scripts into standardized, accessible APIs.
- Containerization: I created Dockerfiles to ensure consistent environments across different stages.
- CI/CD Pipelines: I used GitHub Actions to automate testing, listing, and deployment triggers.
- Cloud Hosting: I explored both AWS EC2 and Railway to deploy services, evaluating scalability and monitoring tools.

Implementation of Demand Sensing One of the core projects I worked on involved demand sensing for supply chain optimization:

- Feature Generation: I worked on separating input data into temporal and non-temporal segments.
- Model Training: I used Optuna for hyperparameter tuning to improve forecasting models.
- Inference and Evaluation: I deployed the models in Docker containers and tested their outputs.
- Scalability Planning: I studied Kubernetes-based orchestration for future scalability.

Tools and Technologies Throughout the internship, I used and became proficient in:

- Python & FastAPI for backend development
- Docker & Docker Compose for container management
- Git & GitHub Actions for automation and collaboration
- AWS and Railway for cloud deployment
- PyTest for automated testing
- Kubernetes for container orchestration (exploratory phase)

Results

- I successfully helped convert and deploy six AI services.
- CI/CD pipelines significantly reduced manual overhead and improved iteration speed.
- I contributed to internal documentation, including setup guides and API references.
- The work allowed me to apply and extend my understanding of ML deployment pipelines.

This experience made it clear how vital infrastructure, collaboration, and scalability are in AI product development. I saw first-hand how Agile practices and DevOps tools fit into real-world workflows, and how theory from machine learning translates into enterprise systems.

The internship was a pivotal step in my professional development. It gave me deep, hands-on exposure to building AI systems that are scalable, maintainable, and useful in business settings. I came away with technical skills, practical insights, and a better understanding of how to deliver AI solutions end-to-end.

REFERENCES

1. Makkar, S., G. N. R. Devi, Solanki, V., "Applications of Machine Learning Techniques in Supply Chain Optimization," Springer, Singapore, ISBN 978-981-13-8460-8, DOI:10.1007/978-981-13-8461-5_98, pp. 965–974, 2020.
2. Prabhudesai, K., et al., "Using Machine Learning and Demand Sensing to Enhance Short-Term Forecasting for CPGs," SAS Global Forum, Washington D.C., USA, Paper No. 4730-2020, April 2020.
3. Mitra, A., et al., "A Comparative Study of Demand Forecasting Models...," Operations Research Forum, Springer, India, ISSN: 2673-8244, Vol. 3, Issue 4, Article No. 58, pp. 1–17, Oct. 2022. DOI:10.1007/s43069-022-00166-4.
4. Docker Documentation. Retrieved from https://docs.docker.com/manuals/
5. Kubernetes Documentation. Retrieved from https://kubernetes.io/docs/home/
6. MIT Sloan Management Review. Retrieved from https://sloanreview.mit.edu/