IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.57
Volume 4, Issue 1, July 2024

Optimized Deployment of Multi-Objective Machine Learning Models in Azure ML: A Compliance-Driven and Cost-Conscious Pipeline Framework

Dheerendra Yaganti
Software Developer, Astir Services LLC, Frisco, Texas.
dheerendra.ygt@gmail.com

Abstract: *The growing complexity of enterprise-grade machine learning (ML) applications demands deployment pipelines that balance performance, compliance, and cost-efficiency. This paper presents a novel framework for the optimized deployment of multi-objective ML models using Azure Machine Learning (Azure ML). The proposed system integrates model evaluation metrics such as prediction accuracy, inference latency, and regulatory compliance scoring to enable intelligent deployment decisions. A cost-aware pipeline is constructed using Azure Pipelines, enabling conditional model promotion across development, staging, and production environments. Compliance alignment is validated through Azure Policy, integrated into the ML workflow to enforce data residency, algorithm transparency, and audit-readiness. Key components of the system include MLflow for experiment tracking, Azure Kubernetes Service (AKS) for scalable inference, and Azure Cost Management for continuous cost analysis. Furthermore, the pipeline incorporates model interpretability tools and automated drift detection to maintain deployment integrity over time. Through rigorous experimentation on classification and regression tasks, the framework demonstrates improvements in deployment efficiency, governance adherence, and cloud resource utilization. This paper offers a reproducible blueprint for organizations aiming to implement secure, cost-effective, and regulation-compliant ML model deployment in cloud-native environments.*

Keywords: Azure Machine Learning, Multi-objective Optimization, Model Deployment Pipeline, Regulatory Compliance, MLflow, Azure Kubernetes Service (AKS), Model Interpretability, Drift Detection, Azure DevOps, Cloud-Native ML, Compliance Scoring, Resource Optimization, Inference Latency

DOI: 10.48175/IJARSCT-19200A