Sreenivas PN

+91-9738078648 | sreenivas.workmail@gmail.com

https://in.linkedin.com/in/srynyvas | Bengaluru, India

PROFILE SUMMARY

Seasoned Machine Learning Engineer with **over 7 years** of experience designing and deploying scalable Machine Learning and Generative AI solutions across enterprise environments. Adept at developing robust ML architectures and hosting LLMs on-premises using GPU-optimized infrastructure. Demonstrated success in implementing Zero Trust security frameworks and building resilient cloud-native solutions on Kubernetes (AKS).

Skilled in automating end-to-end pipelines using CI/CD tools and promoting API-first development strategies. Deep expertise in operationalizing AI models to support real-time inference and analytics at scale. Focused on accelerating GenAI adoption across organisations while ensuring reliability, security, and performance. Experienced in aligning AI capabilities with business objectives to deliver quantifiable impact. Passionate about engineering reusable, production-grade AI systems that drive innovation.

SKILLS

- Project Management
- Azure
- MLOps
- GitHub Actions
- Model Lifecycle Management
- Agile Methodology
- Microservices (Rest API)
- CI/CD/CT
- Deep Learning
- Release Management
- Argo CD
- Azure Pipelines
- Large Language Models
- Machine Learning
- Gen AI
- Software Architecture & Solution Design
- Model Inferencing Servers
- Python
- Event Driven Architecture
- Kubernetes
- Statistical Data Analysis
- On-Prem Model Hosting
- Docker

WORK EXPERIENCE

Jun 2023 – Present

Senior Machine Learning Engineer | Bosch Global Software Technologies, Bengaluru

- Spearheaded design and development of robust internal Software Development Platform to accelerate Generative AI adoption across Bosch's automotive business units, fostering enterprise-scale innovation.
- Architected comprehensive end-to-end platform solution, overseeing program planning, product release cycles, and ensuring timely, high-quality delivery aligned with business and stakeholder requirements.
- Directed deployment of virtual GPU partitioning and NVIDIA Multi-Instance GPU (MIG) technology on H100 GPUs across five Azure Kubernetes Service (AKS) clusters, enabling secure, isolated, and efficient LLM inference environments.
- Delivered scalable, production-grade LLM hosting infrastructure supporting 100+ concurrent enterprise users, with focus on performance, reliability, and data privacy.
- Applied Zero Trust security principles to safeguard platform access, aligning with Bosch's enterprise cybersecurity policies and compliance frameworks.
- Developed and launched developer portals via API Management (APIM), driving an API-first strategy that streamlined integration workflows, governance, and developer onboarding.
- Automated CI/CD processes across 30+ GitHub repositories using GitHub Actions, implementing multi-branch strategies, custom quality gates, and environment-specific load testing to ensure secure, reliable, and auditable deployments.
- Enabled strategic decision-making at executive level by delivering platform adoption analytics, contributing to $12M in realized value 40% of Bosch's annual $30M target in Q1 alone.
- Led platform engineering initiatives and infrastructure planning, managing high-impact team of 7 developers to deliver secure, scalable, and reusable GenAI capabilities across organisation.
- Facilitated architecture design sessions and technical demos with internal stakeholders to onboard federated use cases, while implementing secure authentication flows and compliant interface integrations

Aug 2022 – May 2023 | MLOps Engineer | Fractal Analytics, Bengaluru

- Engineered Self-Service Deployment (SSD) framework for Data Science team to simplify and accelerate ML model onboarding and operational workflows.
- Streamlined deployment lifecycle by integrating GitLab with Argo CD, allowing automated promotion of models into Kubernetes environments with minimal manual involvement.
- Enabled data scientists to independently manage model deployment and lifecycle, reducing dependency on DevOps teams and significantly boosting deployment velocity.
- Implemented shadow deployment techniques to replicate production traffic into UAT environments, supporting safe evaluation of new model versions and code changes. This improved release reliability, minimised rollback risk, and enhanced pre-release test coverage.
- Established an automated model re-training system using Jenkins pipelines and Kubernetes CronJobs, supporting scheduled, hands-free retraining workflows to ensure continuous model relevance and performance in live environments.

Apr 2019 – Aug 2022

R&D Engineer | ABB Global, Bengaluru, India

- Designed and developed microservices-based analytics application using Flask to support diverse business requirements across multiple process industry domains.
- Built automated data preprocessing pipelines to transform raw data from data lake into model-ready formats, streamlining model development lifecycle.
- Created dynamic, generic platform capable of consuming real-time streaming data, training various machine learning models, and autonomously selecting most suitable algorithm based on data characteristics and problem context complete with descriptive statistics and training performance reports.
- Enabled process optimisation and intelligent user recommendations by leveraging insights derived from trained machine learning models.
- Developed data visualisation tool to automate generation of comprehensive reports for wide range of process data scenarios.
- Integrated MLflow to manage complete lifecycle of machine learning models, including version control, automatic retraining, and experiment tracking.
- Utilised model registry services (Azure ML / MLflow) to securely publish models for real-time predictions and classification tasks.
- Built CI/CD pipelines to containerise applications using Docker and deploy seamlessly across on-premises and Azure-based Kubernetes clusters.
- Engineered specialised application for automated time series analysis, enabling training and forecasting with both univariate and multivariate industrial process data.
- Implemented advanced time series methodologies, including exponential smoothing, ARIMA for univariate models, and VAR for multivariate forecasting, to predict future trends and key performance metrics.

INTERNSHIP

Duration	Designation	Organisation
Aug 2018 – Apr 2019	Data Science Intern	ABB Global, Bengaluru

EDUCATION

- Jain University, Master of Computer applications (MCA) – Nov 2024
- IIsc, Bangalore, PG Certification DataScience & MLOps – Dec 2023
- Manipal Academy Of Higher Education, PG Diploma in DataScience – Jul 2018