

Henok Sahilu Mengistu, PhD

henok.sahilu2006@gmail.com
github.com/henokyen
www.linkedin.com/in/HenokMengistu
720-492-3111 USA Citizen

Summary

A results-driven Machine Learning and MLOps Engineer with 8 years of experience delivering large-scale models and optimizing data infrastructure to solve complex business challenges. Proven track record in reducing costs, accelerating delivery, and driving innovation through automation and cross-functional collaboration. Expert in seamless model deployment and continuous improvement in dynamic, fast-paced environments.

Technical Skills

- **ML Expertise :** Transformers, NLP, CNNs, RNNs, LSTM, Unsupervised/Self-Supervised Learning, Computer Vision, Large Language Models (LLMs) application development and deployment
- **Programming:** Skilled in Python, C++, software Engineering Methodologies, and Algorithm Design.
- **Data:** SQL, PySpark, Snowflake, data preprocessing, dimensionality reduction, and feature engineering
- **LLM:** RAG, BERT, OpenAI, Hugging Face models, Elasticsearch, Langchain, and Ollama
- **Mathematics:** Strong foundation in linear algebra, calculus, and probability theory.
- **Machine Learning Tools:** PyTorch, TensorFlow, Keras, Pandas, Opencv, and Scikit-learn
- **MLOps:** Docker, Kubernetes, GithubActions, Jenkins, Airflow, Mlflow, Databricks, and Sagemaker

Work Experience

Sunrun Senior Machine Learning Engineer Jul 2020 - Present

- Designed and implemented an instant response generator using a RAG system for customer feedback, boosting customer satisfaction by 27%. Architected a scalable and real-time inference infrastructure using Amazon SageMaker and pre-mounted EFS volumes, ensuring high availability and efficient resource utilization
- Worked closely with the product team to improve CSAT scores by 23% through the implementation of GenAI customer feedback analysis utilizing BERTopic and llama3.3, achieving a 12% reduction in customer response time.
- Collaborated with the engineering team to deploy the GenAI customer feedback analysis workflow using Docker, AWS EventBridge, and SageMaker, ensuring efficient and scalable production deployment.
- Worked with product to develop an XGBoost model for 15 minutes solar production forecasting, outperforming baselines by 9% (RMSE) and 7.5% (MAE), significantly improving customer satisfaction.
- Partnered with engineering to cut 400 monthly agent hours in solar installation planning using U-Net and graph network model for roof footprint inference, enhancing operational efficiency.
- Developed an MLOps platform leveraging GitHub Actions and AWS SageMaker to streamline and scale the training and deployment of Computer Vision models.

Accenture Machine Learning Engineer Nov 2016 - Jul 2020

- Developed and deployed a tree-based model for USPS package arrival prediction, reducing customer response time by 25%, and productionized it as a REST API on AWS using Docker and Flask.
- Optimized Accenture's internal ChatBot with topic modeling, enriching the knowledge base and increasing weekly traffic by 75%.
- Implemented a CNN for Schlumberger's technician tracking system, improving hard-hat detection accuracy from 95% to 99.96% through data augmentation and regularization techniques.

Education

- Ph.D. Computer Science, University of Wyoming
- M.Sc. Computer Science, Addis Ababa University
- B.Sc. Statistics and Computer Science, Addis Ababa University

Certifications

- AWS Certified Solutions Architect - Associate (ZW6ELTD2H21QQ9WY)
- AWS Certified Machine Learning - Specialty (7LVFT9613NQ1Q7CX)