Eswar Sai Korrapati
Open To Relocate | 551-344-5356 | eksai0726@gmail.com | LinkedIn | Github | Portfolio

Professional Summary

AI/ML Engineer with 3+ years of experience designing and deploying scalable ML solutions on AWS using SageMaker, MLflow, and Spark. Skilled in building end-to-end pipelines, integrating APIs, and applying MLOps practices for reproducibility and automation. Experienced in model explainability with SHAP/LIME and optimization for real-time inference in production environments.

Professional Experience

Software Engineer – AI/ML
Comcast, San Jose, CA
Aug 2024 – Present

- Designed and deployed production-grade ML models using TensorFlow and PyTorch, improving prediction accuracy by 20% in real-time streaming systems.
- Built reproducible ML workflows using Apache Airflow, MLflow, and DVC, enabling versioned pipelines and consistent experiment tracking.
- Deployed models using AWS SageMaker and Lambda, cutting infrastructure costs by 15% and reducing deployment time by 35%.
- Applied SHAP and LIME to create explainability dashboards, aiding stakeholders in data-driven decision-making and regulatory compliance.
- Automated hyperparameter tuning using SageMaker Autopilot and Optuna, leading to an 18% improvement in model F1 score.
- Built scalable batch inference pipelines with Apache Spark and AWS S3, reducing inference time by 25% across large datasets.
- Integrated LLM-based chat capabilities using Bedrock API and prompt engineering, enhancing customer support response automation.
- Collaborated in Agile teams using Git, Docker, and Jenkins to streamline CI/CD across ML development lifecycles.

Stack: Python, TensorFlow, PyTorch, MLflow, SageMaker, DVC, Airflow, Spark, AWS (S3, Lambda, CloudWatch, Bedrock, Sagemaker), Docker, Git, Optuna, SHAP, LIME

Software Engineer
Dentsu, Hyderabad, India
May 2020 – Jul 2022

- Engineered RESTful APIs and backend logic using Django and Flask, increasing system throughput and API response efficiency.
- Developed automated ML workflows using Apache Airflow, improving pipeline reliability and scheduling across data ingestion tasks.
- Containerized and deployed web services via Docker and AWS EC2, achieving high availability and faster deployment cycles.
- Built a centralized Feature Store to standardize feature reuse across models, improving experimentation speed and consistency.
- Leveraged AutoML tools for ad campaign forecasting, reducing manual tuning time by 40% and improving prediction accuracy.
- Integrated model evaluation metrics (AUC, Precision, Recall, F1 Score) and visualization tools like Tableau, improving stakeholder reporting efficiency and clarity.
- Managed cloud resources using AWS, automated deployments with Jenkins, and versioned projects via Git and GitHub Actions, leading to more reliable and faster software releases.

Stack: Python, Flask, Django, SQL, MongoDB, AutoML, Airflow, Feature Store, Docker, Jenkins, Git, AWS EC2, Tableau

Education

Masters in Computer Science
Montclair State University, NJ
Graduated: May 2024 | GPA: 3.65

Projects

- RAG-based PDF Summarizer | Link
- Next Word Prediction with BERT Transformer | Link
- Machine Transliteration | Link
- AI Meeting Preparation Agent | Link
- Job Application Automation | Link

Technical Skills

Programming Languages: Python, C++, SQL, JavaScript
AI/ML & NLP: TensorFlow, PyTorch, Scikit-learn, Keras, BERT, Transformers, GPT, Hugging Face, LangChain, LangGraph
MLOps & Deployment: MLflow, SageMaker, DVC, Airflow, Docker, Kubernetes, CI/CD, Feature Store, Jenkins
Explainability & Evaluation: SHAP, LIME, AUC, Precision, Recall, F1 Score, Confusion Matrix
Cloud & DevOps: AWS (S3, EC2, Lambda, SageMaker, CloudWatch, Bedrock), GitHub Actions
Data & Processing: Apache Spark, Pandas, NumPy, MongoDB, MySQL, Tableau
LLMs & APIs: OpenAI API, Prompt Engineering, LangChain, FastAPI, REST APIs, RAG,LLLMOps,
Tools & Collaboration: Git, Jupyter, Streamlit, Agile, IntelliJ, VS Code
Core CS: Data Structures and Algorithms