Shashank Reddy Kandimalla

• kandimallashashankreddy@gmail.com • Linkedin • GitHub

EDUCATION

San Jose State University (San Jose, USA) Jan 2023 - Dec 2024
Master of Applied Data Science GPA: 3.6/4
Sri Indu College of Engineering and Technology (Hyderabad, IN) Aug 2016 - Sep 2020
Bachelor of Technology - Electronics and Communications CGPA: 8.9/10

SKILLS

- **AI ML:** Large Language Models, Computer Vision, Time Series Analysis, RAG Systems, Fine-tuning, Traditional ML,H20
- **Data Engineering:** Databricks, Spark, Dataflow, Feature Engineering, Data Privacy, Model Evaluation,Sql
- **Development:** Python, TensorFlow, PyTorch, Docker, Kubernetes, Agile Methodologies
- **Cloud Infrastructure:** AWS, Azure, GCP, Containerization, Version Control, CI/CD
- **Business Analytics:** Cross-functional Collaboration, Technical Documentation, Project Management

CERTIFICATIONS

- Google Cloud Certified: **Professional Machine Learning Engineer**

WORK EXPERIENCE

Astranetix Inc **San Jose, US**
AI Engineer Intern Aug 2024 - Present

- Led development of enterprise **Multimodal RAG** system resulting in 60% faster query responses and 45% improved user satisfaction. Implemented custom validation protocols reducing errors by 30% while managing cross-functional collaboration.
- Developed and optimized vector store architecture achieving 55% accuracy improvement and $5K cost savings. Led technical presentations and established best practices for enterprise-scale deployments.
- Architected metadata validation framework improving response accuracy to 90% across enterprise deployments. Implemented **Redis** caching(TTL,LRU) reducing latency by 60% while maintaining sub-second performance at scale.
- Designed a real-time monitoring system with automated recovery protocols, achieving a 40% reliability improvement. Integrated AsyncOpenAI for handling large-scale data processing while maintaining consistent performance metrics

Flatirons AI **San Jose, US**
AI Product Engineer Intern May 2024 - Aug 2024

- Spearheaded development of NLP pipeline integrating Microsoft **GraphRAG**, achieving 50% latency reduction. Led end-to-end project management and implementation of performance optimization strategies.
- Implemented agent-based learning system handling complex financial queries, achieving 40% accuracy improvement. Developed automated validation framework and fine-tuned using the ORPO and SFT+DPO for enhanced model performance.
- Architected microservices deployment handling 100K+ daily records with 99.9% uptime. Established CI/CD pipelines using Docker and GitHub Actions, implementing robust error recovery mechanisms.

Infosys Ltd **Hyderabad, India**
Data Scientist Feb 2021 - Dec 2022

- Architected scalable data pipelines using Hadoop and Airflow for complex preprocessing and feature engineering, implementing a feature store that enhanced model training efficiency and improved predictive accuracy for legal outcomes by 35% across distributed systems.
- Developed and deployed multiple machine learning models in parallel using Prophet, LSTM, and LightGBM for time series forecasting and outcome prediction, achieving 25% performance improvement through systematic A/B testing and hyperparameter optimization.
- Mentored junior data scientists in best practices for model development and deployment, while establishing comprehensive documentation and knowledge sharing processes that reduced onboarding time by 40% for new team members.
- Built production-ready monitoring dashboards and APIs for batch inference, implementing asynchronous operations that improved system performance by 30% while maintaining clear communication of complex technical implementations to non-technical stakeholders.
- Led cross-functional initiatives to detect and address data quality issues using Databricks, implementing automated drift detection algorithms that maintained 95% model accuracy and 97% reliability for subscription services while coordinating with engineering teams to optimize data pipelines.

ACADEMIC PROJECTS

- Developed a reinforcement learning framework for code generation using **BiLSTM**, achieving 90% accuracy through comprehensive validation and testing pipelines, ensuring robust performance and CI for model improvements.
- Modeled an attention-based reasoning system leveraging **ResNet101** architecture for complex inference tasks, enhancing user satisfaction through effective dialog management in real-time processing environments.
- Created a semantic analysis platform using **BERT** and **XGBoost**, maintaining high accuracy through automated monitoring, systematic validation, and implementing robust error handling mechanisms effectively.