Arun Kumar Chukkala

AI/ML Engineer | Arunkiran721@gmail.com | (409) 549-3003 | USA |
LinkedIn | GitHub | Portfolio Website | Why Hire Me

Summary

Experienced AI/ML Engineer with 3+ years of strong background in developing and deploying machine learning models to solve complex business problems. Proficient in utilizing advanced algorithms, deep learning, and AI techniques to drive insights and optimize performance. Skilled in cloud platforms like AWS and Azure for scalable solutions, along with expertise in data processing, model training, and hyperparameter tuning. Adept at working with large datasets, ensuring high-quality performance, and delivering impactful results that improve decision-making, resource allocation, and operational efficiency.

Technical Skills

- **Machine Learning:** XGBoost, LightGBM, CatBoost, Isolation Forest, Autoencoders, LSTM, BERT
- **Data Processing & Engineering:** Scikit-learn, TensorFlow, Pandas, Numpy, Apache Kafka, Apache Spark
- **Cloud Platforms:** AWS (S3, SageMaker, Lambda), Azure (Data Factory, Machine Learning, Monitor)
- **Hyperparameter Tuning:** Optuna, Bayesian Optimization, Grid Search, Randomized Search
- **Programming Languages:** Python, SQL
- **Version Control & CI/CD:** Git, GitHub, Docker, Kubernetes
- **Tools & Frameworks:** Jupyter, TensorFlow, PyTorch, Azure Machine Learning
- **Deployment & Monitoring:** AWS Lambda, Azure Monitor, Real-time performance tracking

Professional Experience

AI/ML Engineer — Jefferies Group
Remote, USA — *03/2024 – Present*

- Designed and deployed advanced machine learning models to predict Customer Lifetime Value (CLV) using over 15 million customer records, integrating data from CRM, transactional systems, and customer interaction logs, driving a 30% improvement in client segmentation accuracy.
- Leveraged AWS services such as AWS S3 for data storage and AWS SageMaker for model training and deployment, processing over 10TB of customer data monthly, and ensuring high scalability and performance in a cloud-native environment with a 99.9% uptime.
- Applied cutting-edge ML algorithms, including XGBoost, LightGBM, and CatBoost, which improved CLV forecasting accuracy by 25%, aiding Jefferies allocate resources more effectively, prioritize high-value clients, and improve customer retention by 18%.
- Led hyperparameter tuning using optimization frameworks like Optuna and Bayesian Optimization, reducing model training time by 40% and improving prediction performance, leading to a 15% increase in predictive accuracy for long-term CLV forecasts.
- Incorporated AI-driven deep learning models like LSTM and transformer-based models such as BERT for customer behavior prediction, capturing complex, long-term engagement patterns and improving customer retention rates by 20%.
- Deployed the CLV prediction model on AWS Lambda for serverless execution, integrating it with Jefferies' microservices architecture, enabling real-time, on-demand insights with zero downtime, resulting in a 25% increase in marketing ROI and improved customer engagement.

Associate AI/ML Engineer — Experian
Hyderabad, India — *01/2021 – 12/2022*

- Developed and implemented a robust anomaly detection system for real-time fraud detection, leveraging machine learning algorithms. Collaborated with a cross-functional team of 10+ members, integrating the solution into Experian's data infrastructure, monitoring over 1 million transactions daily for suspicious activity.
- Leveraged advanced AI techniques, including deep learning and reinforcement learning, improving fraud detection accuracy by 30% and adapting the system to identify emerging fraud patterns, handling over 500,000 unique fraud cases annually.
- Built data pipelines to collect and preprocess large volumes of transactional data, handling over 10 terabytes of data monthly. Utilized Azure Data Factory to ensure seamless data integration, significantly reducing data processing time by 25%.

- Applied anomaly detection methods like Isolation Forest and Autoencoders to detect outliers and potential fraud patterns within financial transaction data, reducing false positives by 40% and improving the system's fraud detection recall by 20%.
- Tuned hyperparameters of machine learning models using grid search and randomized search techniques, achieving a 15% improvement in model performance, which resulted in a 10% increase in fraud detection accuracy while reducing false positives.
- Utilized Python (Scikit-learn, TensorFlow, and Pandas) for data processing, model development, and performance evaluation. Deployed models using Azure Machine Learning, optimizing scalability and processing 2M transactions daily with 98% uptime.
- Watched the performance of deployed models in production, retraining them with updated transaction data every quarter. Ensured model adaptation, cutting undetected fraud incidents using Azure Monitor for performance tracking and real-time alerts.

Education

Master of Science (M.S.) in Computer Science — Lamar University
Beaumont, USA — *01/2023 – 12/2024*

Bachelor of Technology (B.Tech.) in Computer Science — Jawaharlal Nehru University
Hyderabad, India — *08/2016 – 05/2021*

Significant Projects

LLM-Powered Code Analysis System github.com/arun3676/llm-code-analyzer

- Developed a multi-LLM code analysis system integrating GPT-4, Claude, and Gemini, increasing bug detection accuracy by 40%.
- Implemented adaptive prompt engineering techniques reducing false-positive code errors by 30%.
- Designed for enterprise deployment with scalable architecture supporting distributed teams.

AI Learning Path Generator with RAG github.com/arun3676/ai-learning-path-generator

- Created a personalized AI learning system using Retrieval-Augmented Generation (RAG), reducing memory usage by 45%.
- Implemented ChromaDB-based vector storage for efficient information retrieval and knowledge management.

Multimodal AI for Medical Diagnosis github.com/arun3676/multimodal-medical-diagnosis

- Built a diagnostic support system combining vision models with LLMs, achieving 92% accuracy in preliminary diagnoses.
- Optimized AI processing for real-time medical decision support with sub-2-second response times.