Durga Katreddi

Master of Science in Data Science, University of North Texas (3.8 GPA)
+1 917-400-7205 | katreddisrisaidurga@gmail.com | linkedin.com/in/sri-sai-durga-katreddi- | github.com/KATREDDIDURGA

SUMMARY

- Innovative AI Engineer and Data Scientist with 5+ years of experience architecting and implementing cutting-edge AI solutions that drive substantial business value.
- Demonstrated expertise in developing and deploying advanced machine learning systems, including LLM-powered applications, predictive analytics models, and RAG pipelines that have reduced operational costs by 22% and generated $12M in annual savings.
- Skilled in full AI development lifecycle from research to production, combining deep technical knowledge of generative AI, neural networks, and MLOps with strong business acumen.
- Proven track record of transforming siloed enterprise data into actionable intelligence through AI-enhanced dashboards and automated workflows, enabling data-driven decision making and significant improvements in efficiency.
- Adept at leveraging diverse technologies including PyTorch, TensorFlow, AWS, and Kubernetes to build scalable, production-grade AI systems that solve complex business challenges.

SKILLS

- **Machine Learning & AI**: Generative AI (VAEs, GANs, Diffusers, Stable Diffusion), Large Language Models (GPT-4, LLaMA, Mistral, Falcon), Retrieval-Augmented Generation (RAG), Prompt Engineering, Fine-Tuning (LoRA, QLoRA, PEFT), AI Agents (AutoGen, CrewAI, BabyAGI), Neural Networks, Computer Vision, Natural Language Processing, Reinforcement Learning, Model Evaluation & Optimization
- **Data Science & Analytics**: Predictive Modeling, Time-Series Forecasting, Anomaly Detection, Clustering (K-Means, Hierarchical), Classification (Logistic Regression, Random Forest, XGBoost), Ensemble Methods, Statistical Analysis, Customer Segmentation, Risk Assessment, Portfolio Analytics
- **Programming & Development**: Python, SQL, NoSQL, Java, PyTorch, TensorFlow, Keras, Hugging Face Transformers, LangChain, LlamaIndex, Flask, Django, Streamlit, Docker, Kubernetes, MLOps, Web Applications
- **Big Data & Cloud**: Hadoop, Spark (PySpark), AWS (S3, Redshift, EMR, EC2), Azure, Vector Databases (FAISS, Pinecone, ChromaDB), MongoDB, ETL Pipeline Optimization, Cloud-Based Solutions
- **AI Ethics & Safety**: Fairness Assessment, Model Interpretability (InterpretML), Adversarial Robustness, Bias Detection, Responsible AI Development
- **Data Visualization**: Tableau, Matplotlib, Seaborn, Power BI, Plotly, Interactive Dashboards, Executive Reporting
- **Software Engineering**: MLOps, CI/CD Pipelines, API Integration, Version Control (Git), Agile Methodologies, Scalable System Architecture, Model Deployment, FastAPI, Flask
- **Optimization Tools**: DeepSpeed, vLLM, Flash Attention, Triton, Optuna, Ray Tune, Hyperopt
- **Cross-Functional Collaboration**: Stakeholder Communication, Business Strategy, Technical Documentation, Process Improvement, Enterprise Integration

EXPERIENCE

Bank of America January 2024 – Present
Software Developer AI *Texas, USA*

- Implemented LangChain-based AI agents for automating KYC workflows, reducing manual verification processes by 70% while ensuring regulatory compliance
- Developed multi-agent systems using Hugging Face Transformers for document analysis, incorporating state-of-the-art vision models for accurate identity verification
- Engineered privacy-preserving biometric verification systems that reduced onboarding time by 65% while complying with global data protection regulations

- Integrated LangGraph for orchestrating stateful multi-agent systems that handle complex KYC decision flows with explainable AI capabilities
- Built RAG (Retrieval Augmented Generation) frameworks using Chroma vector database to provide contextual regulatory knowledge to AI systems
- Implemented comprehensive LLMOps monitoring with Weights & Biases to ensure model performance and security in production
- Designed sandboxing mechanisms for AI agents to prevent security vulnerabilities while allowing automated processing
- Deployed static machine learning models with proper governance to prevent automated retraining without human oversight, meeting regulatory requirements

University of North Texas August 2023 – December 2023

Student Assistant *Texas, USA*

- Developed a demonstration project leveraging GPT-3.5 and CLIP to create an interactive learning tool that helped students visualize relationships between text prompts and generated images.
- Built and maintained a benchmark testing environment for comparing LLaMA 2, Mistral 7B, and Falcon model performances, which became a core resource for the department's AI curriculum.
- Helped professor design and implement a practical course project where students fine-tuned Stable Diffusion models on domain-specific datasets, resulting in 5 exceptional student projects.
- Created and presented a technical seminar on the architecture and capabilities of GitHub Copilot and CodeLlama, demonstrating practical applications to professors.
- Presented findings from student projects at the department's monthly internal seminar, highlighting innovative applications of Whisper and AudioLM for speech recognition tasks.
- Assisted students in troubleshooting and optimizing their implementations of multimodal models including GPT-4V and ImageBind, improving project completion rate from 70% to 95%.
- AI Ethics & Safety – Ensured fairness and robustness with tools like AI Fairness 360, InterpretML, Adversarial Robustness Toolbox.
- Retrieval-Augmented Generation (RAG) – Built AI systems integrating FAISS, Pinecone, ChromaDB, LangChain, LlamaIndex for improved retrieval.
- LLM Evaluation & Optimization – Benchmarked and optimized LLMs using DeepSpeed, vLLM, Flash Attention, Triton, reducing inference costs.
- Collaborated with professor to develop a comparative analysis of DALL-E 2 and Midjourney outputs, which was incorporated into the department's introduction to AI ethics curriculum.

Cognizant(American Express) December 2020 – December 2022

Software Engineer with Machine Learning *India*

- Developed and deployed an XGBoost model to predict default risk for small business credit card holders, influencing $6M+ in pre-tax income gains and enhancing risk management strategies.
- Improved default prediction accuracy by 5% using a stacking ensemble model, achieving a 150-basis-point Gini lift and capturing more high-risk defaulters through focused analysis of unstructured data using Python and Scikit-learn.
- Productionized a default risk model for 1.5M+ monthly cardholders, enabling real-time decision-making for credit limits, case setups, and promotional offers using AWS and Flask.
- Architected a scalable modeling pipeline using Hadoop and AWS, optimizing ETL workflows and reducing model training time by 30%.
- Engineered customer segmentation models using K-Means and Hierarchical Clustering, improving targeted marketing strategies and increasing customer retention by 15%.
- Built classification models (Logistic Regression, Random Forest, XGBoost) to predict customer churn and default probabilities, achieving 88% accuracy and providing stakeholders with actionable insights.
- Designed interactive dashboards in Tableau and Matplotlib to visualize default risk trends, customer churn, and portfolio performance, improving executive decision-making.
- Automated reporting workflows, reducing manual effort by 10+ hours per month and enhancing operational efficiency through continuous integration (CI) practices and effective version control strategies using Git.
- Leveraged NLP techniques to analyze customer sentiment, driving a 15% improvement in customer satisfaction scores through proactive insights.
- Presented data-driven insights to senior leadership, aligning AI-driven risk assessments with business objectives and contributing to literature reviews on AI methodologies.

UJR Corporate Solutions Pvt. Ltd. August 2019 – November 2020
Data/Decision Scientist *India*

- Engineered a 100-day historical data processing pipeline in MongoDB, implementing profile-based imputation that averaged three years of historical patterns to ensure 30% improvement in data completeness and quality.
- Developed sophisticated anomaly detection methods to identify outliers across large-scale financial datasets, reducing forecasting errors by 25% while ensuring compliance with financial standards using Python and NumPy.
- Created specialized machine learning models including ARIMA, LSTM, CNN, and XGBoost to capture different market behaviors across 3 market segments using TensorFlow and scikit-learn.
- Implemented an ensemble model architecture that dynamically weighted predictions based on market conditions, improving forecast accuracy by 15% for critical financial indicators.
- Implemented an ensemble model to aggregate predictions from 24 models, increasing forecast accuracy by 15%, leading to more precise financial and market predictions using XGBoost.
- Optimized MongoDB query performance through compound indexing and aggregation pipelines, reducing data retrieval times by 60% and enabling real-time dashboard visualizations.
- Designed a RESTful Flask API with stateless architecture for seamless dashboard integration, reducing manual intervention by 40% and supporting horizontal scaling as user traffic increased.
- Implemented comprehensive error handling and logging systems that increased pipeline reliability from 92% to 99.7% completion rate for daily processing tasks.

PROJECTS

AI System for Financial Anomaly Detection

- Built an end-to-end anomaly detection system using Python, FastAPI, and PyTorch, deployed on Azure for real-time performance monitoring.
- Integrated NoSQL (MongoDB) and SQL databases for data storage, enabling scalable analytics for structured and unstructured financial datasets.
- Implemented robust CI/CD pipelines and containerized services using Docker and Kubernetes.

AI Powered Contract Reader:

- Developed an AI-powered contract reading and summarization system leveraging LLMs (GPT-4, Claude), Retrieval-Augmented Generation (RAG), and LangChain, enabling efficient extraction and summarization of key legal terms from complex contracts.
- Integrated vector databases (FAISS/ChromaDB) to store and retrieve document embeddings, improving the speed and accuracy of legal document analysis while reducing manual review time by over 50

AI-Powered Pronunciation Feedback Generator using Whisper and DeepSeek:

- Developed an AI-based pronunciation practice system using Whisper for speech-to-text processing and DeepSeek for language modeling, providing real-time feedback on pronunciation accuracy to enhance fluency.
- Integrated AI Agents to analyze pronunciation patterns, delivering personalized suggestions for improvement and enabling a more interactive and effective language learning experience.

AI-Powered Health Checkup:

- Developed a Streamlit-based Medical Report Assistant that extracts text and tables from PDFs using pdfplumber, generates personalized medical suggestions using LangChain (integrating ChatOpenAI, FAISS, and HuggingFaceEmbeddings), and implements secure user authentication via session state management.
- Leveraged Python, OpenAI GPT-4, PySpark, Pandas, and AWS to streamline the extraction, analysis, and recommendation process, improving efficiency in generating health insights from medical reports.

Alternative Health Recommendation RAG Model:

- Developed a Streamlit-based Health Chatbot that integrates FAISS retrieval, Hugging Face embeddings, and Ollama's LLaMA3.1 model to provide medical insights from novel, published, peer-reviewed research publications.

Wild Fire Prediction:

- Predicted wildfire hotspots using LSTM, Attention, and Transformer models, improving forecast accuracy by 20

Brain Tumor Image Classification:

- Developed a CNN-based brain tumor detection model using TensorFlow, Keras, OpenCV, NumPy, and Scikit-learn.

Tableau - Sales Insights Dashboard:

- Designed and implemented an interactive Tableau dashboard analyzing 93.62M in revenue across 13 Indian markets, showcasing expertise in data visualization, SQL database integration, sales metrics tracking, market segmentation analysis, and customer revenue profiling.
- Identified top-performing products and key accounts to optimize sales strategies and business decisions.