Durga Katreddi

+1 917-400-7205 | katreddisrisaidurga@gmail.com | linkedin.com/in/sri-sai-durga-katreddi- | github.com/KATREDDIDURGA

SUMMARY

AI/ML Engineer with 5+ years of experience building production-grade AI solutions in finance, credit risk, fraud detection, and compliance automation. Proven track record delivering $6M+ in measurable business impact through predictive modeling, customer segmentation, and AI-driven workflow optimization. Skilled in Python, SQL, Java, Go, C++, C#, and large-scale data platforms (Snowflake, Hadoop, Spark). Hands-on expertise in advanced AI/ML frameworks (PyTorch, TensorFlow, XGBoost, LangChain, LangGraph, Semantic Kernel, Hugging Face Transformers) and GenAI technologies including RAG, multi-agent orchestration, and LLM fine-tuning. Strong background in deploying AI systems on cloud platforms (AWS, GCP, Azure) with MLOps and DevOps best practices. Experienced in AI Agent Orchestration, Azure AI Services, Azure Automation, Azure DevOps, and infrastructure optimization/tuning. Known for translating complex data into actionable insights through intelligent dashboards, automated workflows, and agentic AI solutions.

SKILLS

- **Programming Languages & Scripting:** Python, Java, Go, C++, JavaScript, TypeScript, SQL, HTML, Node.js
- **AI/ML Frameworks:** PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, LangChain, LangGraph, Semantic Kernel, Hugging Face Transformers, OpenAI APIs
- **Multi-Agent & Advanced AI:** AI Agent Orchestration, RAG, LLMOps, Multi-Agent Systems, Agent Traceability, Prompt Engineering, PEFT/LoRA, Whisper, CLIP, Stable Diffusion
- **Cloud & DevOps:** AWS (EC2, S3, EMR, Lambda), GCP (Vertex AI), Azure (Azure AI Services, Azure Automation), Kubernetes, Docker, Azure DevOps, Databricks, CI/CD, Infrastructure as Code
- **Databases & Data Processing:** SQL Databases, NoSQL (MongoDB), Snowflake, Hadoop, Hive, Spark, ChromaDB, FAISS, Pinecone, LlamaIndex, ETL Pipelines, BigQuery, Data Lakes
- **Visualization & Reporting:** Tableau, Power BI, Streamlit, Matplotlib, Seaborn, Interactive Dashboards
- **MLOps & Infrastructure:** Model Deployment, Drift Detection, Performance Monitoring, A/B Testing, Weights & Biases, Infrastructure Optimization & Tuning, Sandboxing AI Agents
- **Domain Expertise:** Credit Risk Modeling, Fraud Detection, Customer Segmentation, KYC Workflow Automation, Document Processing, Regulatory Compliance

EXPERIENCE

Bank of America **January 2024 – Present**
AI Software Developer

- Implemented GenAI-powered LangChain agents to automate KYC workflows, reducing processing time by 70% and enhancing compliance efficiency.
- Built multi-agent KYC automation system processing 50,000+ documents daily with full decision traceability and audit trails.
- Designed stateful AI agent orchestration with LangGraph & Semantic Kernel, enabling autonomous planning, execution, and refinement of compliance workflows.
- Developed RAG frameworks using ChromaDB & FAISS serving 10K+ daily queries with 99.9% uptime and <200ms latency.
- Deployed production LLMs with PEFT/LoRA fine-tuning, achieving 92% accuracy improvement for KYC document classification.
- Optimized infrastructure performance using Azure DevOps pipelines & Azure Automation, reducing deployment times by 40%.
- Architected observability dashboards with Weights & Biases to detect drift, hallucinations, and performance degradation in real-time.
- Engineered secure LLM APIs with role-based access, key rotation, and partner integrations.
- Built Kafka-based pipelines for inter-agent messaging, scaling to 100K+ messages/hour with guaranteed delivery.
- Deployed containerized AI services with Docker & Kubernetes across AWS & Azure, achieving 99.95% availability.

University of North Texas **August 2023 – December 2023**
Student Assistant

- 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 using PyTorch and Hugging Face Transformers, which became a core resource for the department's AI curriculum.
- Created and presented a technical seminar on SAS programming for analytics, demonstrating practical applications for financial data analysis to students.
- Developed Snowflake data pipelines and SQL optimization techniques for handling large datasets, improving query performance by 40% for student research projects.
- Assisted students in troubleshooting and optimizing their implementations of Python and SQL code for data analysis projects, improving project completion rate from 70% to 95%.
- Collaborated with professor to develop a comparative analysis of credit scoring methodologies using XGBoost and LightGBM, which was incorporated into the department's financial analytics curriculum.
- Implemented advanced data visualization techniques using Matplotlib and Seaborn to create compelling presentations for academic conferences and student projects.
- Mentored junior developers and students in modern development practices, including Git version control, CI/CD pipelines, and containerization with Docker.
- Developed automated grading systems using Python and machine learning algorithms, reducing manual grading time by 60% while maintaining accuracy.
- Created comprehensive documentation and tutorials for AI/ML model development, serving as reference materials for over 200 students in the program.
- Implemented A/B testing frameworks for evaluating different teaching methodologies and learning outcomes in data science courses.
- Built RESTful APIs using FastAPI and Flask for student project submissions and automated evaluation systems.
- Integrated cloud services (AWS, GCP) into the curriculum, providing hands-on experience with modern deployment practices for machine learning models.
- Developed cross-platform applications using JavaScript and TypeScript for educational tools, enhancing the learning experience for data science students.
- Conducted workshops on prompt engineering and LLM fine-tuning techniques, preparing students for careers in AI and machine learning.

Cognizant(American Express) **December 2020 – December 2022**
Software Engineer with Machine Learning

- 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 with multiple algorithms (Random Forest, LightGBM, XGBoost), achieving a 150-basis-point Gini lift and capturing more high-risk defaulters.
- Productionized a default risk model for 1.5M+ monthly cardholders using AWS cloud infrastructure, enabling real-time decision-making for credit limits, case setups, and promotional offers.
- Architected a scalable modeling pipeline using Hadoop, Snowflake, and AWS, optimizing ETL workflows and reducing model training time by 30% through parallel processing.
- Engineered customer segmentation models using K-Means and Hierarchical Clustering with scikit-learn, improving targeted marketing strategies and increasing customer retention by 15%.
- Built classification models (Logistic Regression, Random Forest, XGBoost) using PyTorch and TensorFlow to predict customer churn and default probabilities, achieving 88% accuracy.
- Designed interactive dashboards in Tableau, Power BI, and Matplotlib to visualize default risk trends, customer churn, and credit portfolio performance for executive decision-making.
- Automated reporting workflows using SAS, Python, and CI/CD pipelines, reducing manual effort by 10+ hours per month and enhancing operational efficiency.
- Leveraged NLP techniques and sentiment analysis using advanced text processing libraries to analyze customer feedback channels, driving a 15% improvement in customer satisfaction scores.
- Presented data-driven insights to senior leadership, aligning AI-driven risk assessments with business objectives and contributing to literature reviews on AI methodologies in lending.
- Implemented MLOps best practices using Kubernetes and Docker for model deployment, ensuring scalability and reliability across production environments.
- Developed comprehensive feature engineering pipelines using Pandas and NumPy, creating over 200 predictive features from raw transaction data.
- Built real-time fraud detection systems using streaming data processing with Apache Spark, reducing false positive rates by 25% while maintaining security standards.

- Created automated model monitoring and alerting systems using Python and cloud services, ensuring model performance consistency and early detection of model drift.
- Integrated multiple data sources using advanced SQL queries and NoSQL databases (MongoDB), creating a unified data warehouse for comprehensive customer analytics.

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

- 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.
- Developed sophisticated anomaly detection methods using statistical analysis and machine learning to identify outliers across large-scale financial datasets, reducing forecasting errors by 25%.
- Created specialized machine learning models including ARIMA, LSTM, CNN, and XGBoost using TensorFlow and PyTorch to capture different market behaviors across 3 market segments.
- Wrote complex SQL queries and stored procedures for data extraction and transformation from multiple sources, creating a unified analytical framework for financial modeling.
- Implemented an ensemble model using advanced stacking techniques to aggregate predictions from 24 models, increasing forecast accuracy by 15% for financial and market predictions.
- 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.
- Implemented comprehensive error handling and logging systems using Python that increased pipeline reliability from 92% to 99.7% completion rate for daily processing tasks.
- Developed automated data quality assessment tools using statistical methods and machine learning algorithms, ensuring data integrity across all analytical processes.
- Built interactive visualization dashboards using Tableau and Matplotlib for stakeholder reporting, enabling data-driven decision making across multiple business units.
- Implemented version control and CI/CD practices using Git and automated testing frameworks, ensuring code quality and deployment reliability.
- Created comprehensive documentation and knowledge transfer materials for data science workflows, facilitating team collaboration and onboarding of new team members.
- Utilized cloud computing resources (AWS EC2, S3) for scalable data processing and model training, optimizing costs while maintaining performance standards.
- Developed time series forecasting models using advanced statistical techniques and deep learning approaches, achieving superior prediction accuracy for financial markets.
- Implemented A/B testing frameworks for model evaluation and business strategy optimization, providing statistical significance testing for decision-making processes.

PROJECTS

FinGuard Agents : Multi-Agent GenAI System for Real-Time Fraud Prediction & Explanation
- Architected an agentic AI system using open-source LLMs, multi-agent workflows, and custom rule engines to predict and explain fraud in financial transactions in real-time.
- Integrated retrieval-augmented generation (RAG) and narrative generation agents to produce compliant, human-readable fraud justifications, improving auditability and user trust.
- Designed agent coordination logic with tool usage control, intent routing, and fallback behavior logging, supporting extensibility across retail, fintech, and regulatory domains.
- Built an event-driven pipeline with modular hooks for LLMOps, graph-based tracing, and domain-specific knowledge bases using Python, LangGraph, and SQLite.

AgentScope : AI Agent Traceability & Decision Debugging Framework
- Designed a general-purpose decision traceability system for AI agents using FastAPI, SQLAlchemy, SQLite, and Streamlit to visualize multi-step decision paths.
- Implemented a backend API to fetch agent run metadata and step-wise execution (thoughts, tools, actions, final response), improving model interpretability, debuggability, and fallback analysis.
- Created a professional UI for agent trace search using Streamlit, enhancing trace audits for LLM, rule-based, and intent-based AI agents across fraud, legal, and retail domains.
- Engineered the system with extensible data models and modular hooks for LangChain, Rasa, and multi-agent orchestration, supporting future LLMOps, compliance, and real-time agent monitoring needs.

VisionAlign: Feedback-Driven Image Generation via Low-Resolution Preview

- Developed a two-stage GenAI image pipeline that first generates low-resolution previews on CPU to align with user intent before committing to full-resolution GPU rendering.
- Minimized compute costs by up to 80% using conditional generation feedback loops and progressive refinement, avoiding redundant high-resolution model calls.
- Implemented a natural language feedback module to capture semantic mismatches (e.g., object pose, context) between user expectations and generated previews.
- Tech stack included Stable Diffusion, CPU-inference U-Net, custom feedback handler, and prompt alteration agent for interactive fine-tuning.

AI Insurance Claims Processing System: NLP & Fraud Detection

- Built an end-to-end AI-powered insurance claims processing system using GPT-4 and advanced NLP techniques, achieving 95% reduction in processing time and 92% accuracy in fraud detection, resulting in $500K annual savings per 1000 claims processed.
- Developed natural language claim processing, automated fraud analysis with confidence scoring, and professional report generation capabilities, creating a modular architecture with real-time analytics dashboards for executive decision-making.