

# Pavan Sabnaveesu

 832 274-8181 |  [sabnaveesuprasanth@gmail.com](mailto:sabnaveesuprasanth@gmail.com) |  GitHub |  LinkedIn

## Professional Summary






AI Software Developer with 5+ years of experience building scalable AI/ML solutions across finance, e-commerce, and healthcare domains. Proficient in Vertex AI, Agents, OpenCV, MLflow, RESTful APIs, GraphSAGE, reinforcement learning, and quantized LLMs, driving latency reduction and performance improvement. Adept at transforming complex data pipelines into high-impact models that elevate decision-making, customer experience, and risk mitigation.

## Education

Texas A&M University  
Master of Science  
Graduate Research Assistant

**CGPA: 3.9/4.0**  
January 2023 – Dec 2024  
February 2023 – Dec 2024

## Technical Skills

 **Programming & Database:** Python, R, SQL, PostgreSQL, DSA, Pinecone, REST& Fast APIs, Flask, Node.js, Reat.js  
 **Visualization & Cloud:** Tableau, Power BI, Matplotlib, Seaborn, Plotly, AWS, Azure, GCP  
 **Machine Learning & NLP:** NumPy, Pandas, Scikit-learn, TensorFlow, Keras, PyTorch, OpenCV, spaCy, NLTK  
 **DevOps & LLMs:** Docker, Kubernetes, CI/CD (Jenkins), Git, OpenAI-GPT, Hugging Face, BERT, RAG  
 **ML algorithms:** Regression, XGBoost, Random Forest, LSTM, CNN ,GRU, Transformer, YOLO, LLaMA

## Work Experience

### Lead Data Scientist, Cyber Nirvana (Contract)

**February 2025 – Present**

- Developed a conversational AI system using LangChain, RAG, and OpenAI GPT, integrating multimodal LLM with model context protocol for real-time financial query processing, achieving 95.3% response accuracy
- Optimized customer support chatbot by 24% using LoRA and QLoRA under PEFT, fine-tuning attention heads and activations while quantizing model layers to reduce latency and GPU memory by 42%
- Developed Azure AI agents for real-time patient triage, integrating quantization with BERT models on Azure ML, improving diagnosis by 28% and cutting time by 25% for a telehealth platform serving 8,000 daily users

### Data Scientist, NEXT ROW Private Limited

**July 2021 – Dec 2022**

- Translated Chinese-to-English translation and speech models using NLTK, Wubi, GRU encoder-decoder, and MFCCs. Deployed quantized pipelines on Vertex AI with RESTful APIs and MLflow for scalable real-time language learning
- Developed ARIMA and quantized LSTM models using PySpark and Kafka on transaction streams, capturing seasonality and market trends. Achieved 20.3% improvement in forecasting accuracy, optimizing retail banking risk
- Implemented a deep Q-learning model using real-time GPS data to minimize route deviations, which increased ETA precision by 17% and saved an estimated \$3 million in operational costs
- Collaborated with cross functional teams to design, build, and deploy ML models for various domains, ensuring robust solutions to business problems

### Data Scientist, Meslova Systems Private Limited

**Sept 2018 - June 2021**

- Developed a recommendation engine using collaborative filtering engine with AWS Neptune and GraphSAGE, processing millions of e-commerce interactions to increase user engagement by 20%
- Conducted A/B testing for subscription churn prediction using quantized GBM models, achieving 15% higher accuracy, reducing customer churn by 20%, and boosting retention ROI by 10%
- Streamlined CI/CD processes with Jenkins and Docker, enhancing model accuracy by 20%, and deployed scalable for fraud detection applications using Kubernetes for continuous performance optimization
- Optimized complex SQL queries to improve data retrieval speed for a data analytics platform by indexing critical columns, restructuring queries to minimize joins, and leveraging subqueries
- Resolved bottlenecks and improved model flexibility within AWS Fraud Detector and SageMaker and models for banking transactions by continuous monitoring, validation, and automated updates, reducing false positives by 20%

## Projects

- Developed real-time object and lane detection for self-driving cars, increasing lane accuracy by 21% using U-Net and improved traffic sign and pedestrian tracking with YOLOv8 and ByteTrack
- Designed a RAG model with Hugging Face transformers using Pinecone DB, boosting document search accuracy by 35% and driving a 15% revenue increase through seamless knowledge base integration using Vertex AI