**Pavan Sai Prasanth Sabnaveesu**

**832 274-8181|** [**sabnaveesuprasanth@gmailcom**](mailto:sabnaveesuprasanth@gmail.com) **|** [**GitHub**](https://github.com/prasanthsabnaveesu) **|** [**LinkedIn**](https://www.linkedin.com/in/prasanth-sabnaveesu/)

**Professional Summary**

Experienced machine learning engineer and developer with over 5 years of expertise in designing, developing, optimizing, and deploying AI/ML models. Proficient in state-of-the-art neural networks, large language models, and real-time applications across domains such as autonomous systems, natural language processing, and recommendation engines. Skilled in end-to-end development, from data preprocessing to deployment, driving measurable business impact through innovative AI solutions

**Education**

*Texas A&M University* Master of Science, Computer Science  ***CGPA: 3.83***

**Technical Skills**

Programming & Databases : Python, R, SQL, PostgreSQL, DSA, Pinecone

Web Development : HTML, CSS, jQuery, Bootstrap, Django, REST API, Fast API, Flask

Data Visualization : Tableau, Power BI, Matplotlib, Seaborn, Plotly

Machine Learning : NumPy, Pandas, Scikit-learn, TensorFlow, Keras, PyTorch, OpenCV

Cloud & DevOps : AWS, Azure, GCP, Docker, Kubernetes, CI/CD (Jenkins)

ML algorithms : Regression, XGBoost, Random Forest, LSTM, CNN ,GRU

Advanced ML : PEFT (LoRA, QLoRA), RAG, GPT, BERT, Transformers

**Work Experience**

***Graduate Research Assistant, Texas A&M University* *Feb 2023 – Dec 2024***

* Designed and deployed a wind turbine blade segmentation system using Mask R-CNN, YOLOv7, and YOLOv8 achieving 98.7% mAP with variable IoU and a 25% reduction in training time
* Developed real-time object and lane detection for self-driving cars, increasing lane accuracy by 21% using U-Net ResNet 101, and improved traffic sign and pedestrian tracking with YOLOv8 and ByteTrack
* Developed CNNs from scratch using NumPy and Pandas, achieving over 99.1% accuracy on image classification
* Researched and implemented state-of-the-art complex neural network architectures and algorithms

**Software Developer – A, NEXT ROW Private Limited** ***July 2021 – Dec 2022***

* Optimized LLM performance by 20% using Parameter-Efficient Fine-Tuning techniques (LoRA, QLoRA)
* 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
* Developed a conversational AI system using LangChain and RAG on AWS, enabling dynamic and accurate responses to user queries across financial and medical domains
* Built speech-to-text systems with MFCCs and GRU/LSTM models, and text-to-speech systems with Tacotron 2 and WaveNet, achieving 85% accuracy.
* Collaborated with cross functional teams to design, build, and deploy AI models for various domains, ensuring robust solutions to business problems
* Enhanced predictive model accuracy by 25% by advanced feature engineering and optimized hyperparameter tuning
* Web scraped from multiple sources, ensuring quality and integrity through thorough preprocessing and validation
* Translated Chinese to English using NLTK and wubi, cleansing and tokenizing sentences for input Employed GRU-based encoder-decoder architecture for accurate language translation

***Software Developer – AI, Meslova Systems Private Limited* *Sept 2018 - June 2021***

* Developed full-stack applications using Python and Django, integrating ML and AI models resulting in a 30% increase in performance and response time
* Optimized and troubleshooted AWS SageMaker model deployments, ensuring seamless performance and adaptability through continuous monitoring, validation, and automated model updates
* Spearheaded the creation of a recommendation system integrated with AWS Neptune graph database, boosting user engagement by 20% through personalized recommendations
* Conducted A/B testing for subscription churn prediction, with XGBoost achieving 15% higher accuracy, reducing 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 ML applications using Kubernetes for continuous performance optimization
* Optimized and developed complex SQL queries to improve data retrieval speed by indexing critical columns, restructuring queries to minimize joins, and leveraging subqueries, enhancing overall application performance