Sunny Savita

Data Scientist & GenAI Engineer

✉ snshrivas3365@gmail.com ☎ +91 8770203258 📍 Bengaluru, Karnataka 📅 10 Nov 1995

Linkedin Github YouTube

PROFILE

With 4 years of experience in Data Science and Generative AI, I have successfully implemented full-stack Computer Vision, NLP, and Generative AI projects using the MLOps Pipeline to make data-driven decisions. I excel in Neural Networks, Transformers, Python, Lang Chain, Llama Index, RAG, Vector Databases, Hugging Face, and building reliable LLM applications

SKILLS

Python: *(Automation | DSA)*, **Linux**, **Statistics**, **Machine Learning**, **Deep Learning:** *(ANN | Pytorch)*, **Computer Vision:** *(CNN | Image Processing | Object Classificaiton | Object Detection | Object Segmentation)*, **Natural Language Processing:** *(RNN | LSTM | Encoder Decroder | Attention |Transfromer | BERT | GPT)*, **Generative AI:** *(Open AI, Mistral , Llama, Gemini, Clude, Embeddings, HuggingFace, RAG, Finetuning, AI Agents)*, **LLM Framework:** *(Langchain | LlamaIndex | AWS Bedrock | Azure AI Studio | Vertex AI)*, **Vector Databases:** *(FAISS | ChromaDB | Pinecone | LanceDB)*, **RestAPIs:** *(FastAPI | Flask)*, **Databases:** *(MySQL | MongoDB | Cassandra)*, **MLOps:** *(GIT | Docker | DVC | MLflow | Kubeflow | GitHub Actions | CircleCI | Terraform)*, **Cloud:** *(AWS | Azure)*

PROFESSIONAL EXPERIENCE

Data Scientist & GenAI Engineer, *iNeuron.ai* ↗ Sep 2020 – Present | Bengaluru, India

- Leading a team of 5+ Data Scientists and Jr. Data Scientists in various Data Science projects
- Introduced and implemented MLOps methodology using MLFlow, DVC, Docker, Kubeflow, and GitHub Actions
- Worked as a R&D member, where key research is focused on drone-based AI solutions, core AI development and generative AI applications.
- Provided AI consultancy services on behalf of iNeuron Intelligence Pvt. Ltd.
- Delivered expert lectures on MLOps, Data Science, and Generative AI to over 1000 students

Deep Learning Intern, *iNeuron.ai* ↗ Feb 2020 – Aug 2020 | Bengaluru, India

INTELLIGENT RADIOLOGIST ASSISTANT (IRA)

- Developed an automatic diagnostic app for radiologists using deep computer vision, tested on Brain Tumor Segmentation MRI data. Analyzed 750 4D volumes and trained segmentation models, using a continuous training pipeline with GitHub Actions on Paperspace and utilized S3 as a model registry.
- Deployed a dockerized app on a GPU instance for faster inferencing.

PROJECTS

Megatron Chatbot, *iNeuron Product* Oct 2023 – present

Tech: Python, LLMs, Whisper, Langchain, Mongodb Vector Search, FastAPI, Docker, AWS S3, EC2, ECR, Github Action, Terraform

Megatron is a full-featured, real-time, RAG-based AI bot that is a part of the Ineuron support system. It can take many forms of input, including text and speech, and was created to lighten the workload of the Ineuron support staff, which is responsible for responding to client inquiries about their products.

- Used in house videos and converstaional data(skype and REVEchat) generated transcript from videos using Whisper model and saved it inside the s3.
- Generate summary using OpenAI API (GPT-3.5) from the transcript and saved inside the s3.
- Created RAG pipeline using summrized data converted into embedding using text-embedding-ada-002 and stored inside the mongo vector search and checked similarity using cosine similarity
- Using langchain with RAG for maintaining meomory of conversation and google search API for searching something outside.
- Used Fastapi as an interface for the prediction
- Implemented CI/CD in monolith architecture using GitHub actions self hosted machine and docker and deployed the app on EC2.

AI based Learning Managment System, Jul 2023 – present
iNeuron Product

Tech Stack: Python, GPT-3.5 turbo, OpenAI text embedding, Lang chain, EC2, ECR, S3, GitHub Actions MongoDB, Docker, Paper Space.

Created an automated system to generate quizzes, assignments, and evolution of assignments for the iNeuron LMS which includes CI/CD for Data Collection, Data Embedding, Model Inferencing and Model Evaluation.

- Data Storage and Accessibility: Utilized Amazon S3 as a central data store, implementing public ACLs for image data to facilitate easy access and listing post-prediction, enhancing data accessibility and management.
- Selected OpenAI text-embedding-ada-002 as a final model for generating embeddings and employed the Annoy algorithm for quick and efficient nearest neighbor searches, achieving logarithmic time complexity for embedding comparisons, which significantly improved the system's performance and response time.
- Implemented NLP Techniques: After multiple experiments selected GPT-3.5 turbo for language modelling tasks, ensuring the accurate generation of quizzes and assignments by using advanced language models for content creation.
- Enhanced Data Pipeline Efficiency: Designed and deployed a CI/CD pipeline using GitHub Actions, ensuring seamless integration and continuous delivery for data collection and model inference processes.

Pipeline Standardization, *iNeuron Internals* Mar 2022 – Jul 2023

- Built a decoupled machine learning system and implemented a project code and deployment structure for all upcoming projects.
- Implemented Google's MLOps level 2 architecture on AWS, Azure, and GCP.
- Introduced MLflow and Weights & Biases for experiment comparison and parameter tracking.
- Drafted MLOps templates for clients and iNeuron internals.

HACKATHONS

Simplified AI, *Online machine Learning Platform* Remote

Tech: SK-learn, Flask, Plotly, Aws, MongoDB, Docker, MongoDB, MySQL.

Solution Build:

- Built a no-code platform enabling users to perform EDA, data preprocessing, feature engineering, model training, process scheduling, and custom scripting.
- Provided one-click solutions for each module without requiring any coding.
- Designed automated and custom training sections, with options for data injection and export to multiple cloud platforms.

COMMUNITY LECTURES

GENERATIVE AI COMMUNITY EVENT

Volunteered for a live Ineuron community session that was published on FreeCodeCamp.org.

COMMUNITY EVENT: GENERATIVE AI END-TO-END PROJECT IMPLEMENTATION,
Enterprise Application: E-Learning App

A playlist explaining how to use AWS services to build and deploy E-Learning Chatbot

MLOPS COMMUNITY EVENT

A community event to guide folks via MLOps.

HONORS AND AWARDS

Star Performer of the Year, *iNeuron*

EDUCATION

B.Tech Computer Science and Engineering, Aug 2015 – Jul 2019 | Bhopal, India
Barkatullah University Institute of Technology

- Cumulative Percentage: 66.0%
- Relevant Coursework : DSA, DBMS, OS, Artifical Intelligence, Data Warehosing & Mining