Gowtham Venkata Sai Ram Maddala

Stony Brook, NY | (934) 246-9689 | gowthamvenkata.maddala@stonybrook.edu | Github | LinkedIn | Portfolio

EDUCATION

Stony Brook University, Stony Brook, New York **Aug 2024 - Dec 2025**
Master of Science in Data Science, ***GPA: 3.84/4***

International Institute of Information Technology, Bangalore, India **Apr 2023 - Dec 2023**
Advanced Certificate Programme in Data Science with Specialization in NLP, ***GPA: 3.8/4***

Koneru Lakshmaiah Education Foundation - KL University, Hyderabad, India **Sept 2017 - May 2021**
Bachelor of Technology in Computer Science with specialization in Data Science, ***GPA: 9.02/10***

TECHNICAL SKILLS

Programming Languages: Python, C, C++, Java, HTML, CSS, JavaScript, R, MATLAB, SQL
Tools and Platforms: Git, Docker, Kubeflow, Flask, FastAPI, Google Cloud Platform (GCP), Microsoft Azure
Frameworks and Libraries: TensorFlow, PyTorch, Keras, NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, CUDA
Machine Learning: Predictive Modeling, GBDTs, Random Forests, Clustering, LLMs (LLaMA, GPT, Mistral)

EXPERIENCE

Soroco **Feb 2024 - Jul 2024**
Software Engineer (Machine Learning) — **Azure DevOps, Python, Docker, Pytest, SQL, APIs** *Bangalore, India*

- Developed 4 Flask APIs to generate flowgraphs of user activities based on screens used for the Workgraph product.
- Trained **YOLOv9** on annotated screenshots to enable the model to accurately detect interacted fields on user's screens.
- Leveraged the **Guidance** library to format **LLM** outputs, reducing post-processing needs, and optimized **Mistral-7B and LLaMA2** models, cutting average inference time from **5 seconds to 0.8** seconds without compromising quality.
- Designed the test cases using **PyTest**, ensuring smooth functionality of **APIs**, and validated JSON outputs using Postman.
- Boosted company **revenue by 15%** through successful onboarding of **3 Fortune 500 clients** in a timeframe of **4 months**.

Awone.ai — Client: Carelon Global Solutions **Apr 2023 - Feb 2024**
Data Scientist — **Python, Bert, DBSCAN, Predictive Modelling, Tensorflow, PCA** *Hyderabad, India*

- Developed a Jira ticket analysis using **BERT and DBSCAN** for effective clustering of ticket summaries and descriptions.
- Converted these text into 200-dimensional vectors using BERT and applied **DBSCAN clustering** on the embeddings.
- Predicted resolution time using **polynomial regression** with **Lasso** regularization, achieving an R-squared value of **0.89**.
- Automated workforce assignment and resource allocation based on cluster analysis, **reducing operational costs by 30%**.

Ivy Comptech **Aug 2021 - Feb 2022**
Software Engineer — **MySQL, Java** *Hyderabad, India*

- Managed a high-volume transactional database with over **3 million** records as part of the wallet/payments team.
- Optimized 30 complex SQL queries, reducing execution time by **30%** and significantly boosting data pipeline performance.

Telescope (Voxlogic.inc) — Acquired by Meta **Jul 2020 - Dec 2020**
Software Development Intern (AI Platform Team) — **Python, TensorFlow, APIs** *Sunnyvale, USA - Remote*

- Architected a conversational search solution using the **TAPAS model** from Hugging Face, enabling numerical question answering on tabular data extracted through a custom web scraping pipeline, achieving **97.45%** accuracy.
- Integrated the model into Slack, allowing users to input tabular data and receive real-time responses, and quantized the model using **TensorFlow** to optimize for speed, ensuring swift responses during conversational searches.
- Contributed to the development of Telescope, which was later acquired by Meta for **$2.4 million** in 2021.

PROJECTS

Ola Driver Churn — GitHub

- Engineered a high-performance driver churn prediction model for Ola using **XGBoost**, achieving **0.97 precision, recall, and 0.98 AUC score**, by analyzing key features such as income, total business value, and quarterly ratings.
- Conducted feature analysis and engineering on Ola driver data, implementing **KNN-based imputation and target encoding** to optimize model performance and identify primary factors influencing driver retention.

Summarizing news articles using LSTMS — GitHub

- Implemented a many-to-many LSTM model to generate summaries from **60,000 news articles**, improving performance by stacking additional LSTMs and applying dropout, achieving a ROUGE-L score of **0.72**.