Sandeep Dwarampudi

+1 (312)-409-1311 | reddydvvsn@gmail.com | LinkedIn | GitHub | Chicago, IL

Data Scientist with 3 years of experience in Data Engineering, ML and GenAI.

Education

Master of Science in Computer Science (Thesis Track) - *University of Illinois Chicago | Chicago, IL*
Relevant Coursework: Machine Learning, Data Processing and Text Mining, Relational Databases, Data Science, Computer Algorithms, Geometric Data Structures, Biomedical NLP

Bachelor of Technology in ECE (Minor in Data Science) - *Manipal Institute of Technology | India*

Professional Experience

NLP Researcher - *UIC Computer Science | Chicago, IL, USA* ***Sept 2023 – Jul 2024***

- Designed a **model architecture using BART** small language model with adapters for biomedical lay summarization, enhancing performance through domain-specific **pre-training on GCP Vertex AI**.
- **Fine-tuned** state-of-the-art GenAI models like **GPT-3.5 Turbo, Gemini 1.5** and **RAG**-based systems for task-specific datasets.
- Achieved **better performance** compared to **open-source LLMs** such as **Gemma2** and **BioMistral**, despite fewer parameters and a smaller model size due to parameter-efficient tuning.

Data Science Engineer - *UIC Pharmacy | Chicago, IL, USA* ***Oct 2022 – Aug 2023***

- Developed a fast-binning algorithm that generates denoised spectra achieving a **50x speed improvement** and deployed it on a **Spark cluster**.
- Created a streamlined data pipeline to automate all the tasks involving large data collection to post-processing using multiprocessing, achieving a **25% reduction in processing time**.
- Extracted high quality data by conducting **extensive statistical analysis** on large datasets and enhanced database efficiency by **optimizing SQL queries**, reducing query execution time by 27% on average.

Data Scientist - *Merkle | Bangalore, India* ***Jul 2021 – May 2022***

- Developed an **NLP-based intent classification** system using the **BERT model** to automatically categorize user queries as part of a query priority ranking system.
- Trained and deployed ML models using decision trees on **AWS SageMaker** to identify high-value users during marketing campaigns, leading to a **20% increase in revenue over 8 months**.
- Designed and executed **A/B testing** using **Adobe Target** to optimize the placement of a banner for a new website feature, achieving a 10% lift in click-through rate and a 5% increase in conversion rate.

Data Scientist Intern - *Merkle | Bangalore, India* ***Mar 2021 – Jul 2021***

- Revamped **uplift models** by implementing T-Learner and S-Learner approaches and tracked their performance using **MLFLow,** resulting in a **15% increase in response rates** by more accurately targeting individuals likely to respond to marketing emails.

Projects

BioLexicon

- Created an ensemble model using BERT, ALBERT and ROBERTa for finding the lexical complexity of a word in a sentence.
- Used weighted layer pooling to efficiently utilize transformer embeddings, leading in a notable 13% enhancement in accuracy.

Wikipedia Continual Learning with RAG

- Utilized data from Wikipedia to build a Retriever-Aware Generation (RAG) model, with Llama-3 model for generating human-like responses and SentenceTransformers with Chroma pre-trained model for semantic similarity search.
- Managed data ingestion, vectorization and storage using Langchain, enabling efficient handling of the data and model interaction.

Real-Estate Data Pipeline using AWS

- Designed and implemented an automated ETL pipeline using Zillow API, AWS (S3, Lambda, Redshift) and Airflow for seamless data flow from extraction to visualization in Amazon QuickSight.

Technical Skills

Programming/Visualization: Python, R, SQL, HTML, CSS, JS, C++, Linux, Tableau, Excel, PowerPoint

Machine Learning:	Clustering, Regression Analysis, Decision Trees, Random Forests, XGBoost, Deep Learning, NLP, LLM, RAG
Big Data/Database:	PySpark, SparkSQL, DataBricks, Airflow, MongoDB, MySQL, SQLite, Big Query, Hadoop, Adobe Analytics
Cloud/MLOps:	Docker, MLFlow, Flask, FastAPI, CI/CD pipelines, Git, AWS SageMaker, EC2, S3, Lambda, EventBridge, RedShift, QuickSight, GCP Vertex AI
Data Science Experience:	A/B Testing, Hypothesis testing, Statistical Modeling, Model Deployment
Libraries and frameworks:	Pytorch, NumPy, Pandas, Scipy, Scikit-Learn, Statsmodels, Bokeh, Matplotlib, Seaborn, Spacy, Huggingface transformers, LangChain, LlamaIndex, NLTK, Pytest

Certifications

Coursera - Deep Learning Specialization