Sandeep D

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

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

Education

Master of Science in Computer Science (Thesis Track) - *University of Illinois Chicago | Chicago, IL* ***Aug 2022 – Aug 2024***

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* ***Jul 2017 – Aug 2021***

Professional Experience

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

- Developed a model architecture by integrating adapter modules into BART and implementing a combination of **self-supervised and supervised learning based pre-training mechanisms** for biomedical lay summarization
- Achieved **better performance compared to SOTA open-source LLMs**, including Gemma2 and BioMistral despite having fewer model parameters and a smaller size.

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** and extracted high quality data by conducting extensive statistical analysis.
- Built an interactive **data visualization application** using Python Bokeh to illustrate gene-expression for cells in various organs of the human body, accelerating data-driven decision-making processes into collagen research.
- Set up a **server with NVIDIA GPUs** and implemented a GPU-accelerated CLI for **running inference** on transformer-based proteomics models.

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

- Developed 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** 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.
- Utilized **Adobe Analytics** for data analysis and **SQL** for ad-hoc tasks, delivering actionable insights and supporting business stakeholder needs.
- Achieved a **40% reduction** in report request volume and generation time by consolidating reports from Excel into **Tableau Server**.

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

- Revamped **uplift models** to target individuals likely to respond to marketing emails, resulting in a **15% increase in response rates**.
- Automated the end-to-end modeling process in Python from one single code base, **saving 80 man-hours per month**.

Projects

BioLexicon

- Created an ensemble model comprising fine-tuned BERT, ALBERT and ROBERTa models for finding the lexical complexity of a word in the given Biomedical sentence
- Used weighted layer pooling to efficiently utilize transformer representations, leading in a notable 13% enhancement in accuracy on test data.

Automated Real Estate Data Pipeline

- Designed and implemented an end-to-end Python ETL process utilizing Zillow Rapid API, AWS services (S3, Lambda, Redshift), and Apache Airflow for workflow orchestration.
- Developed automation for data extraction, transformation, and loading, ensuring seamless data flow from source to visualization using Amazon QuickSight.

HR attrition

- Leveraged advanced statistical analytics such as bayesian inferences, latent variable analysis, and implemented a Logistic- XGBoost ensemble model to predict attrition among a dataset of 30,000 employees, ensuring compliance with company policies.
- Generated key insights aimed at reducing the hiring rate by 12%, focusing on factors such as employee satisfaction, career development opportunities, and managerial effectiveness.

Technical Skills

Programming/Visualization:	Python, R, SQL, HTML, Linux, Tableau, Excel, PowerPoint
Machine Learning:	Supervised, Unsupervised, Clustering, NLP, Large Language Models (LLM), RAG, Generative AI
Big Data/Database:	PySpark, ETL Data Pipeline, Airflow, MongoDB, Cassandra, Hadoop, Adobe Analytics
Cloud/MLOps:	AWS (SageMaker, EC2, S3, Lambda, IAM roles), Docker, Flask, CI/CD pipelines, Github
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

Achievements

- Awarded merit-cum-means scholarship for the 4 years of my undergraduate study

Certifications

- Coursera - Deep Learning Specialization