Anup Sarkar

Data Science & Generative Al Practitioner | Former MSSQL DBA

anupsarkar8106@gmail.com +91 8106754538 Hyderabad

Portfolio LinkedIn GitHub Indian

Professional Summary

Data Science and Generative AI practitioner with a strong foundation in database administration and engineering. Bring over three years of proven MSSQL database administration experience, now focused on applying expertise in data cleaning, exploratory data analysis (EDA), feature engineering, and machine learning model development using Python and scikit-learn. Proficient in fine-tuning large language models (LLMs), prompt engineering, and deploying models through Flask APIs. Demonstrated ability to integrate analytical skills with data engineering to deliver scalable, end-to-end data solutions

Certifications

Data Science Certificate
Cedlearn Tech (NASSCOM)
Mar 2025

SQL Server DBA
Udemy
March 2022

Professional Experience

AI & Machine Learning Developer-Mental Health Chatbot, *Elevate_Labs-Intern* 08/2025 – 10/2025
Hyderabad

- Developed an AI-driven mental health chatbot using BlenderBot, fine-tuned on the MentalChat16K dataset to handle emotional and psychological support conversations.
- Designed a full preprocessing pipeline for text cleaning, deduplication, and dataset splitting into train, validation, and test sets to ensure model generalization and robustness.
- Fine-tuned the BlenderBot-400M-distill model on CPU using optimized training loops, reduced epochs, and batch handling for efficient performance on limited compute.
- Incorporated data tokenization, attention masking, and label alignment to enhance conversational accuracy and reduce token truncation errors.
- Evaluated model performance using exact match accuracy, macro precision/recall/F1 scores, and ROUGE-L metrics on validation and test sets.
- Built an inference module enabling live user interaction and integrated it into a FastAPI-based backend for real-time responses.
- Developed a lightweight frontend using HTML, CSS, and JavaScript to simulate web chat interaction with the deployed model.
- Deployed the fine-tuned model to a Hugging Face repository for seamless API-based loading and integration into the web application.
- Tools & Technologies: Python, PyTorch, Hugging Face Transformers, Pandas, Scikit-learn, FastAPI, HTML/CSS/JS, Google Colab

Machine Learning Engineer – Semantic Book Recommendation System, 07/2025 – 08/2025
CedLearn Hyderabad

- Built a semantic book recommender using Kaggle's 7k Books dataset, enriching features with emotion detection and semantic embeddings.
- Applied a DistilRoBERTa-based emotion classifier on book descriptions to quantify affective tone and reader sentiment.
- Generated dense vector embeddings using Sentence-Transformers (MiniLM) and stored them in Chroma DB for efficient vector retrieval.
- Designed a multi-stage ranking pipeline combining semantic similarity, genre classification, and emotion-aware filtering to enhance personalization.
- Developed and deployed an interactive Gradio dashboard visualizing book covers, authors, and summaries for user exploration.
- Referenced the project framework and design methodology from Scaler Academy's *Recommender Systems* learning modules.
- Improved recommendation coherence and emotional relevance through embedding space optimization using cosine similarity metrics.
- Tools & Technologies: Python, Pandas, LangChain, Hugging Face Transformers, Sentence-Transformers, Chroma DB, Gradio.

Machine Learning Developer – Sarcasm Detection (Deep Learning, NLP), 05/2025 – 06/2025
Scaler Hyderabad

- Built a sarcasm classification model on 26K+ headlines using TensorFlow and Keras, achieving 83.05% test accuracy.
- Applied text preprocessing techniques including tokenization, padding, lemmatization, and stopword removal to enhance NLP performance.
- Designed a Sequential Neural Network with Embedding and GlobalAveragePooling1D layers for semantic feature extraction.
- Visualized data patterns using Matplotlib and Seaborn to guide model optimization.
- **Tools & Technologies:** Python, TensorFlow, Keras, NLTK, NumPy, Pandas, Matplotlib, Seaborn.

Machine Learning Developer - Fraud Detection System, *CedLearn-intern* 03/2025 – 04/2025
Hyderabad

- Performed data cleaning, exploratory data analysis (EDA), and preprocessing on approximately 6 million transaction records, including handling missing values and filtering merchant transactions.
- Engineered new features such as transaction hour, transaction day, and merchant destination flag to improve predictive accuracy.
- Applied label encoding for categorical variables and resolved severe class imbalance using oversampling techniques.
- Trained and fine-tuned a Random Forest Classifier, optimizing hyperparameters to enhance model performance.
- Achieved 92% recall for fraud cases, validated using classification reports and performance visualizations.
- **Tools & Technologies:** Python, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn.

Machine Learning Engineer - IVF Outcome Predictor, *CedLearn-intern* 02/2025
Hyderabad

- Developed and implemented a machine learning model to predict IVF pregnancy outcomes using Python and scikit-learn.
- Cleaned and pre-processed clinical and treatment data by handling missing values with forward fill and encoding categorical variables.
- Engineered new features such as age bins and IVF treatment classifiers to improve model performance and interpretability.

- Achieved 88% accuracy with a Random Forest model and optimized its performance using GridSearchCV for hyperparameter tuning.
- **Tools & Technologies:** Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.

Senior Engineer – SQL Server Database Administrator, *Microland LTM* 01/2022 – 11/2024
Bengaluru

- Administered and optimized 150+ SQL Server instances and 1,100+ databases across production and non-production environments, ensuring high availability, security, and optimal performance.
- Automated health checks, maintenance plans, and performance tuning tasks, significantly reducing manual workload and enhancing system reliability.
- Designed and implemented robust backup and recovery strategies, including Always-On Availability Groups and Failover Clustering, achieving near-zero downtime.
- Conducted query tuning, index optimization, and performance analysis to improve application responsiveness and throughput.
- Managed SQL Server version upgrades (2012, 2016, 2022), database migrations, and security configurations to meet compliance and business continuity standards.
- Tools & Technologies: SSMS, SQL Server Agent, T-SQL, SQL Profiler, Extended Events, SQL Configuration Manager, DMVs, Always-On Availability Groups, Failover Cluster Manager

Technical Skills

Programming & Machine Learning — Competent
- **Languages & Libraries**: Python (Pandas, NumPy, scikit-learn, TensorFlow, Keras)
- **Concepts**: Machine Learning, Deep Learning, NLP, Generative AI, LSTM, Prompt Engineering
- **Model Development**: Feature Engineering, EDA, Model Evaluation, LLM Fine-Tuning

Visualization & Business Intelligence — Competent
- **Tools**: Power BI (DAX, Power Query), Tableau
- **Skills**: Dashboard Design, Data Storytelling

Databases & SQL — Proficient
- **Technologies**: MSSQL, MySQL, SSMS, SSIS
- **Expertise**: Performance Tuning, Query Optimization, Stored Procedures

Cloud Platforms & Tools — Amateur
- **Platforms**: AWS, Azure
- **Related Tools**: Azure Data Studio, AWS S3, SQL Workbench, Flask (for Model Deployment)

Education

Bachelor of Engineering (B.E), Computer Science & Engineering, 2016 – 2020
Osmania University Hyderabad