Anshul Arya

+91-9826157271 | anshul.arya1996@gmail.com | LinkedIn

About Me

Data-Driven Scientist with 6 years of expertise in Machine learning, NLP, Generative AI, and Data Engineering.

Professional Summary: Proficient in building AI-driven solutions like automated text classifiers, text generation and chatbots. I am skilled in Python, SQL, and Microsoft Azure, with expertise in data pipelines, advanced analytics, and interactive dashboards. Delivered impactful insights across sectors like Oil & Gas, Pharma, Banking.

Technical Proficiency

- **Programming Languages:** Python, SQL
- **Machine Learning:** Regression Models, Decision Trees, Random Forest, SVM, XGBoost, Ensemble Methods.
- **NLP & Text Processing:** NLTK, TF-IDF, Bag of Words, Word2Vec, RNN, LSTM, Encoder-Decoder Models, Transformers.
- **Libraries & Tools:** NumPy, Pandas, Scikit-learn, Seaborn, Matplotlib, Keras, TensorFlow.
- **Generative AI Skills:** LangChain, Langsmith, LLMs Models, RAG Framework, Hugging face, Ollama, Open AI embeddings, Vector Data base, Streamlit-based Applications
- **Cloud Platforms:** Microsoft Azure (Azure Data Lake Storage and Databricks)
- **Data Engineering:** Experience in building scalable data pipelines, including the implementation of Medallion Architecture (Bronze, Silver, Gold layers) for improved data processing and analytics.
- **Visualization:** Tableau, Streamlit

EXPERIENCE

Centelon Solutions | Bangalore, India (March 2019 - Present)
Senior Consultant - Data Science
Working for the Data team, mainly responsible for suggesting Machine and Deep learning solutions by analysing data.

Projects:

1. **Email Triage | Bizessence Australia | NLP Based Classification Project**
   - **Objective:** Developed an automated email triage system using NLP and Machine Learning to classify emails, improving workflow efficiency and automating email organization
   - **Approach:**
     - **Data Processing:** Pre-processed email text data, including text **normalization, tokenization**, and **feature extraction** (TF-IDF, Bag of Words).
     - **Model Development:** Implemented various binary classification models (**Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost**) to analyse the email content and classify emails into actionable and non-actionable categories.
     - **Automation:** Integrated the final model into an automated pipeline to move classified emails to the respective subfolders in real-time.
   - **Impact:** Improved email management efficiency by automating the classification and organization process, reducing manual sorting time by **98%.**

2. **Gen AI HR Policy Chatbot | RAG Application using Open AI and OLLAMA**
   Tool: Python, Ollama LLMs (Mistral, Gemma2:2b), OpenAI Embedding, RAG Framework, NLP, LangChain, Langsmith, Streamlit
   - **Objective: Developed a chatbot using RAG (Retrieval Augmented Generation) architecture to enable employees to query HR policies seamlessly.**
   - **Approach:**
     - Built an interactive chatbot leveraging **Ollama LLMs (mistral, gemma2:2b)** for dynamic Q&A functionality tailored to HR policy queries.
     - Trained models using company-specific HR policy PDFs, enabling **accurate and secure** responses to employee questions.
     - Implemented a modular architecture using **LangChain** for seamless prompt handling, embedding creation, and response generation
     - Designed a user-friendly **Streamlet** interface and integrated **Langsmith** for API tracking and model performance monitoring.

3. **Data Migration & Visualization | Oil & Gas Company in Australia**

Tool: Microsoft Azure, Databricks, Tableau

- **Objective:** Designed and implemented a data migration and visualization solution using the Medallion Architecture to optimize data processing and enhance data-driven decision-making.
- **Key Contributions:**
  - **Data Pipeline Implementation:** Developed a robust data pipeline using Microsoft Azure, incorporating Medallion Architecture (Bronze, Silver, Gold layers) to ensure data integrity, scalability, and efficient processing.
  - **Interactive Dashboards:** Created dynamic Tableau dashboards that transformed complex data sets into visually compelling insights, enabling stakeholders to effortlessly identify key trends and make informed decisions.
  - **Client Collaboration:** Engaged closely with clients to understand their requirements, translating them into actionable solutions that exceeded expectations.

4. **Document Parsing | Leading Pharma Firm**

Tool: Python, Apache Tika, Tesseract

- **Requirement:** To develop a solution that can extract data from PDF invoices and store it in the database within 15 minutes after receiving mail from vendors.
- **Approach:**
  - Used Tabula and Camelot for tabular data extraction and custom-trained Tesseract for text recognition.
  - Leveraged Apache Tika to handle diverse PDF patterns and extracted metadata from emails (sender ID, subject, date, and attachments).
  - Developed and implemented complete AI pipeline to process PDFs dynamically and store structured data efficiently.

CERTIFICATIONS

- Complete Guide to Building, Deploying, and Optimizing Generative AI with LangChain and Hugging face.
- Post Graduate Program in Big Data Analytics & Optimization, INSOFE.
- Data Engineering Nanodegree, Udacity.
- Qlik Sense for Data Science and BI, Udemy.

ACHIEVEMENTS

- I received a chapter award for excellent work in my current position.
- Nominated for the **Spot Award** four times at Centelon for exceptional performance on client projects.
- Participated in WNS Analytics Wizard 2018 (Machine Learning Hackathon) and secured all India 13th rank.

Classroom Training– Detail

INSOFE, Bangalore - Classroom Training – Data Science (June 2018 – Dec 2018)

Project:

1. **Predict whether a patient will be readmitted within 30 days using patient, hospital, and diagnosis data.**
   - Built classification models (Logistic Regression, Decision Tree, Random Forest) to identify patients likely to readmit within 30 days.

ACADEMIC PROFILE

- **Mandsaur Institute of Technology, RGPV-Bhopal** (Aug 2014 - May 2018)
  B.E (Electronics and Communication Engineering, CGPA – 7.33)
- **Adarsh Malwa H S School** (March 2013 - April 2014)
  Intermediate (M.P Board)