Aaryan Shah

Los Angeles, CA | (213)998-1789 | shah.aar27@gmail.com | LinkedIn: aaryanshah27 | Github: Aaryan04

Education

University of Southern California | Los Angeles, CA — Aug 2023 - May 2025
Master of Science in Applied Data Science — GPA: 3.64/4.0
Coursework: Data Mining, Machine Learning, Database Systems, DS Professional Practicum, UX Design and Strategy

Dwarkadas J. Sanghvi College of Engineering | Mumbai, India — Aug 2019 - June 2023
Bachelor of Technology in Electronics and Telecommunication Engineering — GPA: 3.72/4.0
Coursework: AIML, Deep Learning, Big Data, Algorithms, Image Processing, OOPS, Computer Networks and Security

Experience

USC Marshall School of Business — Dec 2023 - Present
NLP Researcher — Los Angeles, CA
- Developed a scalable data pipeline to process 250,000+ unstructured PDF files on energy sector deals, reducing data ingestion time by 40% through automated text extraction using PyMuPDF and NLP preprocessing
- Built a hybrid RoBERTa model with rule-based scoring, improving strategy classification accuracy by 25%, and evaluating shifts from core to emerging technologies on a 1-5 scale, driving better technology adoption decisions
- Boosted model accuracy by 35% by fine-tuning GPT and BERT models to classify 50k+ unstructured business deals in the oil and gas sector, using domain-specific clustering to identify key trends in mergers and acquisitions

Gustovalley Technovations — Feb 2022 - May 2022
Machine Learning Intern — Remote
- Deployed ML models on cloud platforms, automating predictions and cutting insight delivery time by 30%
- Tuned hyperparameters using grid search and Bayesian optimization, reducing prediction errors by 25%
- Built 25+ PowerBI dashboards, providing actionable insights for strategic decision-making

Technocolabs Softwares — Jun 2021 - Aug 2021
Software Developer Intern — Remote
- Reduced project costs by 15% by implementing an advanced PostgreSQL-based cost tracking system, improving financial analysis and uncovering key savings areas for more efficient resource management
- Improved page load speed by 40% by revising MongoDB queries, dramatically enhancing application performance

Academic Projects

Advanced Martian Frost Detection Using Deep Learning | *Python, Tensorflow, Keras*
- Devised a custom 3-layer CNN with HiRISE data, resulting in 92% training and 82% validation accuracy and trained 25+ epochs with ReLU activation, batch normalization, L2 regularization, and ADAM optimizer
- Improved model performance and adaptability employing data augmentation, regularization, and early stopping
- Leveraged and analyzed VGG16, ResNet50, and EfficientNetB0, demonstrating effectiveness in image classification

YelpRec: Scalable Hybrid Recommendation System | *Python, PySpark, Spark RDD, XGBoost, CatBoost*
- Developed a hybrid recommendation system utilizing item-based collaborative filtering and XGBoost regression on the Yelp dataset, resulting in a 15% increase in prediction accuracy compared to CatBoost regressor
- Built a scalable data pipeline using PySpark and Spark RDDs to process 1 million+ records across 6 JSON datasets, achieving a 25% reduction in processing time for rapid experimentation with multiple recommendation algorithms
- Conducted hyperparameter tuning and model optimization for XGBoost, fine-tuning parameters such as learning rate, tree depth and number of estimators, achieving an RMSE of 0.9798

DocuBot: Intelligent PDF Query System | *Python, LangChain, OpenAI, FAISS, Streamlit, PyPDF2*
- Led the development of a conversational retrieval chain using LangChain's memory buffer and OpenAI's GPT models, which improved answer precision by 70%, delivering highly relevant responses based on user queries
- Optimized document parsing and embedding pipeline using PyPDF2 and LangChain, resolving bottlenecks and implementing a chunk overlap strategy, reducing chunking time by 30% and improving context retention by 50%
- Integrated FAISS vector store for optimized management of text embeddings, achieving an 85% accuracy rate in retrieving relevant document sections during LLM-based query responses

Dynamic Sales Performance and Forecasting Tool | *PowerBI, Excel, DAX*
- Developed a comprehensive PowerBI sales dashboard that tracked over $1.6M in sales and 22K orders, providing leadership with real-time insights to improve decision-making across key regions and customer segments
- Built a 15-day sales forecast model with historical data and DAX, predicting 10.6K units for peak periods, and created 10+ interactive visualizations to drive product promotions and improve inventory planning

Technical Skills

Languages/ Databases: Python, R, C++, MySQL, MongoDB, PostgreSQL, Spark RDD, Hadoop, Firebase, Redis
Libraries: PyTorch, Tensorflow, Keras, LangChain, NLTK, Gensim, SpaCy, Scikit-Learn, Seaborn, Matplotlib
Cloud: AWS (EC2, S3, Athena, Glue, Kinesis, Lambda, Eventbridge, Bedrock), GCP (Cloud Storage, BigQuery)
Tools/Frameworks: Git, Tableau, Power BI, DBT, Apache Kafka, Airflow, Beautifulsoup, Selenium, Jira, Excel

Publications

- "Prediction System Design for Monitoring the Health of Developing Infants from Cardiotocography Using Statistical Machine Learning", published in Design Engineering, Scopus International Journal, Volume 2021, Issue 07 — **link**