RAKSHIT SAKHUJA

+65 91322353 ◇ Singapore
rakshitsakhuja@gmail.com ◇ rakshitsakhuja

OBJECTIVE

Senior Data Scientist with 8.5+ years of expertise in machine learning, recommender systems, LLMs, and Generative AI. Proven track record of building real-time search, ranking models, RAG pipelines, and end-to-end ML systems for enterprise applications. Experienced in cross-functional leadership, client-facing engagements, and production-grade deployments.

EDUCATION

Indian Institute of Technology Hyderabad August 2020 - July 2023
Master of Technology in Data Science
Thesis: Survey on Natural Language Understanding for Tabular Data using TAPAS and TableFormer
Project: Query Auto Completion using Deep Learning

Dr. A.P.J. Abdul Kalam Technical University August 2012 - June 2016
Bachelor of Technology in Computer Science

SKILLS

Languages: Python, R
ML & AI: Generative AI, LLMs, RAG, Fine-Tuning, Recommender Systems, Semantic Search, Model Interpretability (SHAP), MLFlow, Statistical Modeling
LLM & Vector Tools: OpenAI, LlamaIndex, LangChain, Chainlit, Faiss, Ollama, llama.cpp
AWS: Sagemaker, Lambda, API Gateway, EC2
Azure: Search, Cosmos, Redis, Kubernetes, ML
Databases: SQL, Teradata, MongoDB, PostgreSQL, ElasticSearch
Libraries/Frameworks: PyTorch, Pyspark, Databricks, SBERT, scikit-learn, spaCy, Tensorboard, Flask, FastAPI
CI/CD: Git, Gitlab, Docker, Kubernetes, AzureDevops

EXPERIENCE

Senior Data Scientist, Mediacorp Pte Ltd November 2022 – Present
Singapore

- Leading the development of video recommendation models (mewatch), managing a cross-functional team of data scientists and engineers to enhance content discovery across multiple platforms (web, mobile, TV).
- Developed **LTR models** using **LightGBM (LambdaRank)**, **SBERT**, and **CatBoost**, leading to a **40% improvement** in CTR for app recommendations, **15% on the web**, and **30% on TV platforms** compared to baseline models.
- Designed and deployed in-house vector indexing API using **gRPC** and **PyTorch**, reducing embedding retrieval latency and returning similar items for real-time recommendation systems.
- Developed multiple recommendation models, including a **GRU4Rec**-based sequential model, **content-based filtering**, **item-based collaborative filtering**, supporting various widgets across the platform.
- Implemented ML pipeline deployments using cloud services, CI/CD (Azure DevOps), Databricks, and Kubernetes for scalable recommendation services.
- Executed **large-scale A/B testing** to evaluate model performance, ensuring continuous model improvement.
- Leveraged LLMs to generate **genres** and **subgenres** for metadata enhancement using content titles and abstracts.
- Designed GenAI-based SQL generation pipelines for Metric data using GPT-4o and Chainlit, improving internal operational efficiency for business users
- Conducted a technical workshop on **RAGs** using **LlamaIndex**, facilitating cross-team knowledge transfer.

Senior Data Scientist, AgreeYa Solutions(Evalueserve) November 2020 – October 2022
Noida, India

- Built a neural search engine using Elasticsearch for B2B clients, indexing documents and news, with auto-suggested filters for organization and location using Named Entity Recognition.
- Developed a recommendation engine using collaborative filtering for personalized news and document suggestions.
- Implemented autocorrection functionality using **Levenshtein distance** (Peter Norvig's method), **Soundex**, and a custom MLE-based algorithm to select the best correction.
- Built low-latency semantic search features such as related entities and related search queries using **SBERT**, **transformers**, **FAISS**.
- Fine-tuned NER model using **spaCy** and **PyTorch** with **BERT** to extract locations and organizations, enabling entity-specific page generation.
- Developed REST APIs using **Python**, **Flask**, **Gunicorn**, **Nginx**, and **Supervisord** to serve the features within the search engine.

Data Scientist, Grail Insights Aug 2018 – November 2020
Noida, India

- **Demographic Modelling:** Built multi-class classification models to predict demographics for POS data and feature engineered consumer KPIs related to buyer behavior for a retail client [**Python, SQL, sci-kit-learn, NumPy, pandas**]
- **Sentiment Analysis:** For Quality Service Reviews, leveraged Naive Bayes as the baseline model, along with LogisticRegression/KNN/Linear SVM and LSTM(RNN) with mini-batches using Word Embeddings [**Pytorch, sci-kit-learn, LSTM**]
- Built an in-house social media listening tool to preprocess open ends via NLP techniques such as Named Entity Recognition, TopicModeling(LDA), Word Embeddings (word2vec, glove), and Word Clouds using tf-idf to derive customized insights [**Python, spaCy, nltk, NLP**]

ETL Developer, Tata Consultancy Services June 2016 – Aug 2018
Mumbai, India

- Built ETL pipelines and automated workflows using **Informatica, Teradata, Python, Unix, and shell script**; performed EDA to clean and merge data from multiple sources

PROJECTS

Agentic Document Analyzer Bot

- Built an LLM-based Q&A system over arXiv papers using a RAG pipeline with open-source embedding models, FAISS, and LlamaIndex for semantic retrieval and interaction.

Building Domain-Specific Language Models

- Co-authored a LiveProject with Manning Publications titled "Building Domain-Specific Language Models," involving the development of n-gram, token-based LSTM, and character-based RNN models.

CERTIFICATIONS

- deeplearning.ai: ChatGPT Prompt Engineering for Developers Jul 2024
- deeplearning.ai: Building Generative AI Applications with Gradio Mar 2024
- NVIDIA: Building Intelligent Recommender Systems Mar 2022
- NVIDIA: Fundamentals of Accelerated Computing with Python(Numba) Sep 2021
- DeepLearning.ai: NLP with Classification and Vector Spaces Jul 2020