Jaideep Bommidi

jaideepbommidi@gmail.com | Linkedin | GitHub | +1 940-629-6953

- **ML Engineer & Data Scientist** with 5+ years building production systems across fintech, rail analytics, and 3D printing; expert in **Python**, **PyTorch/TensorFlow**, **LLMs/RAG**, **retrieval & ranking**, and **AWS/Azure** deployment.
- Deep expertise in **retrieval systems** (**vector search with FAISS**, embedding models, hybrid retrieval), **LLM fine-tuning (LoRA)**, and ranking optimization (**NDCG/MAP metrics**); plus end-to-end ML ownership from data pipelines (**SQL**, **Spark**, **Snowflake**) and feature engineering to **A/B testing** and low-latency deployment via **REST APIs**, **Docker/Kubernetes**, and **CI/CD**.
- Proven impact: Architected **enterprise LLM agent for agentic AI applications with multi-stage retrieval + re-ranking** for Temenos Journey Manager/Maestro; shipped **self-hosted, bank-grade AI systems** with PII scrubbing and compliance guardrails; delivered churn prediction (+30% retention), defect detection (95% accuracy), and real-time dashboards driving business decisions.

Education

M.S | Advanced Data Analytics | University of North Texas — May 2024 | USA

M.S | Image Processing and Computer Vision | University of Bordeaux — Aug 2018 | France

B.Tech | Electronics and Communication Engineering | National Institute of Technology, Trichy — May 2016 | India

Technical Skills

Generative AI & LLM Tools: GPT, LangChain, BERT, LlamaCpp, OpenAI, Hugging Face, Pinecone, Vector Databases

Programming: Python, R, C/C++, SQL, JavaScript, Shell Scripting.

Machine Learning & AI: Agentic AI, **Scikit-learn**, **TensorFlow**, **PyTorch**, Keras, NLP, classical ML (trees, boosting, clustering, PCA), time series (ARIMA, Prophet, VAR), computer vision, A/B testing, LLMs/agents (prompt engineering, RAG, vector search with **FAISS**).

Web & APIs: **FastAPI**, **Flask**, **Django**, **React.js**, **AngularJS**, REST.

Data Engineering & Big Data: **SQLAlchemy**, **Apache Spark**, **Hadoop/HDFS**, **Hive**, **Pig**, MapReduce, **Snowflake**, **Redshift**, **HBase**.

Databases & Warehouses: **PostgreSQL**, **MySQL**, **SQL Server**, **MongoDB**, NoSQL.

IaC, Containers & Orchestration: **Terraform**, **CloudFormation**, **Docker**, **Kubernetes** (**EKS**, **ECS**).

MLOps & CI/CD: **Git**, GitHub, GitLab, **Jenkins**, AWS **CodeBuild/CodeDeploy/CodePipeline**, GoCD, Maven (gmavenplus).

Testing & QA: **Selenium**, **Cypress**, JUnit, Mocha, Jest, Enzyme.

Visualization & Analytics: **Tableau**, **Power BI**, Microsoft Excel, SAS Enterprise Miner.

Cloud Platforms: **AWS** (S3, EC2, Lambda, RDS, SageMaker), **Azure** (Data Factory, Data Lake), **GCP**.

Experience

Machine Learning Engineer (AI Systems) — May 2025 – Current

Webster Bank — *Denton, TX*

- Led the architectural design and deployment of an e**nterprise-scale LLM agent** for Temenos Journey Manager/Maestro, encompassing **multi-stage retrieval, re-ranking, and semantic search** (FAISS vector store). Defined the overall system architecture for agentic AI applications, including data ingestion pipelines, FAISS vector store selection, scoring heuristics, and integration with existing systems.
- **Embedding-based retrieval pipeline:** ingested internal properties/logs into a **vector store (FAISS indexing)**, built **hybrid search** (dense embeddings + sparse keyword filters) for least-privilege context augmentation, and fine-tuned ranking to improve **NDCG/MAP by 18%** in offline eval.
- **Self-hosted, bank-grade deployment of agentic AI systems:** ran **open-weight LLMs and embedding models (Hugging Face)** on company servers with zero external calls; added PII scrubbing, policy guardrails, and full audit logging to meet security/compliance SLAs.
- Shipped low-latency inference services using **caching, batching, and fallback logic** to maintain <200ms p95 response times; established **monitoring dashboards** for drift detection and guardrail compliance in production.
- Collaborated cross-functionally with security/compliance and platform teams to iterate on LLM/NLP components (scoring strategies, retrieval plugins, tool integrations) and build custom DDS services, form validators, and RBAC features.

Data Scientist — Sep 2024 – May 2025

Sypnios — *Dallas, Texas*

- Built **end-to-end ML pipelines** for segmentation, impact/risk scoring, and forecasting using **PyTorch, scikit-learn, and pandas**; implemented feature engineering, cross-validation, and ran **A/B experiments** tied to user metrics to validate model impact.
- Designed and shipped an **AI agent-powered report generator** using **retrieval-augmented generation (RAG)** over operational/project data, leveraging **Google Cloud Platform services such as Compute Engine and Cloud Storage** for data processing and model deployment. Orchestrated semantic search **(embedding models + vector retrieval)** to fetch relevant documents, then synthesized executive-ready PDF/Excel/Power BI narratives with compliance guardrails.
- Developed analytics-driven web application end-to-end (**Python, FastAPI, SQLAlchemy**) covering data modeling, ETL, and dashboarding; exposed ML models via **REST APIs**, containerized with **Docker**, and deployed to **AWS/Azure** with **CI/CD pipelines and monitoring** for reliability.
- Established **data quality checks, schema validation, and lineage tracking**; optimized SQL for batch/incremental loads to improve SLAs and accuracy, ensuring reproducibility for training and evaluation workflows.
- Created interactive dashboards (**Power BI/Tableau**) and embedded analytics surfacing real-time KPIs for commercial and NGO stakeholders, enabling data-driven decision-making at scale.

Data Scientist (Software Engineer) Mar 22 – Jan 23

German Railways (Deuta Werke GmbH) *Cologne, Germany*

- Performed statistical and ML analysis on incident & performance data for drivers and trains; identified patterns that informed risk mitigation and operational KPIs.
- Built risk/incident prediction pipelines in **Python** (**pandas**, **NumPy**, **scikit-learn**); engineered features (intersections, normalization, label encoding) and validated with cross-validation.
- Developed time-series **fleet demand forecasting models** (ARMA/ARIMA, exponential smoothing) for German railway operations, processing 147k+ rows/hour of real-time train telemetry to optimize fleet utilization, predict maintenance needs, and inform capital planning decisions.
- Implemented streaming ingest and black-box telemetry decoding, automating analysis of **>147k rows/hr** for incident detection and driver behavior scoring.
- Designed and optimized **SQL** for data collection, migration, and augmentation (**SQL Server**, **PostgreSQL**); collaborated with data engineers/ops on ETL design and data quality.
- Ran **A/B testing** to refine incident-reporting rules and reduce false positives; documented assumptions and acceptance criteria.
- Built big-data pipelines across **Hadoop/HDFS** using **Hive**, **Pig**, and **MapReduce** for large-scale transforms and access.
- Led AWS migration with **Terraform**; containerized and deployed services on **EKS/ECS**; leveraged **S3**, **RDS**, **EC2**, and **Lambda** for scalable data processing.
- Exposed analytics via **RESTful APIs** and internal tools (**Django**, **React**); enabled self-serve insights for stakeholders.
- Orchestrated **CI/CD** with **AWS CodeBuild/CodeDeploy/CodePipeline**, plus **Jenkins** and **Docker**; versioned with **Git** in **Jira**/Kanban Agile; added unit tests (**Mocha**, **Enzyme**, **Jest**) and conducted code reviews.
- Delivered dashboards and a reporting repository in **Tableau** (daily/monthly summaries, trends, and benchmarks) for operations and leadership.

Data Scientist Sep 20 – Dec 21

iFactory3D GmbH *Düsseldorf, Germany*

- Built customer segmentation using **DBSCAN**, **K-means/K-means++**, and **Hierarchical** clustering; tailored printer features by segment, driving a **40% increase in customer satisfaction**.
- Developed churn prediction and retention recommender models (**Decision Trees**, **Random Forest**, **Reinforcement Learning**); achieved a **30% lift in retention** and, via **A/B testing**, a **15% increase in sales conversions**.
- Performed market targeting with **factor** and **cluster** analysis for investor reports, contributing to a **20% increase in investment returns**.
- Built sales and campaign-forecasting models (**ARIMA**, **VAR**, **NNAR**) improving forecast accuracy by **25%**.
- Developed a deep-learning defect detector for 3D prints using **TensorFlow** with **CUDA** acceleration and **transfer learning**; reached up to **95% accuracy** and **+20%** performance over baseline.
- Deployed the full model pipeline on **AWS** (**S3**, **EC2**, **Lambda**); implemented a lightweight **C++** inference module on the printer motherboard, reducing processing time by **40%**.
- Designed **SQL/NoSQL** schemas and data-integrity checks, cutting data errors by **99%**; accelerated retrieval via **Hive** on **Hadoop/HDFS** and **Redshift**, reducing query times by **50%**.
- Built internal data apps and APIs (**REST**), integrating **AngularJS** front end with **Spring** services on AWS; improved app performance and UX by **35%**.
- Operationalized ML with **Docker** and CI/CD (**Jenkins**, **AWS CodeBuild/CodeDeploy/CodePipeline**); versioned with **Git** in an **Agile** workflow; conducted reviews and mentored interns.

Software Developer Jul 16 – Aug 18

Temenos India Pvt Ltd. *Full-time, Chennai, India*

- Developed and provided support for modules within the Temenos product.
- Collaborated with the EB team to design and implement optimized bank transactions using JDBC, SOAP, and Tomcat, resulting in substantial transaction speed improvements.
- Conducted thorough research aimed at designing and optimizing SVM algorithms to extract valuable insights from images.

Research Internships

Graduate Teaching Assistant Aug 23 – May 24

University of North Texas *Denton, TX*

- Instructed and supported courses in **Machine Learning**, **Big Data**, and **Tableau** for **100+** students; provided technical guidance and fostered a strong learning environment.
- Developed comprehensive course materials—programming assignments and projects—using **Python**, **R**, **Hadoop**, and **Tableau**.
- Built a Python-based web scraper to collect submission metadata; anonymized and analyzed data with **pandas** and **NumPy**; applied time-series methods (**ARIMA**, **Exponential Smoothing**, **Prophet**) to uncover study-behavior patterns.
- Engineered ETL pipelines for multi-source integration, leveraging **SQL**, **Hadoop**, **HBase**, **SQL Server**, and **PostgreSQL** to ensure reliable data flow and quality.
- Developed and deployed deep learning models with **TensorFlow** and **PyTorch** for the UNT Swim Team: image segmentation and keypoint detection to derive stroke angles/velocities; integrated with physiological metrics (e.g., heart rate, personal best times) to support performance analysis.

Data Scientist Intern Feb 20 – Jul 20

Solarvibes GmBH *Master Thesis, Berlin, Germany*

- Developed a deep learning plant disease classification algorithm using computer vision techniques such as image segmentation and classification.
- Engineered a cloud architecture capable of handling 10,000 simultaneous requests with dynamic load balancing for scalability on AWS.
- **European Space Imaging Project:** Developed advanced machine learning models to analyze farmland images retrieved from AWS S3, utilizing Geographic Information Systems (GIS) for comprehensive vegetation analysis. Implemented AI-enabled data-driven models to enhance predictive accuracy in agricultural monitoring.
- Engineered a robust cloud architecture on AWS to support real-time data processing and analysis, ensuring scalability and efficient resource management. Leveraged AWS services such as Amazon SageMaker for model training and deployment, and Amazon Kinesis for streaming data analytics.
- Collaborated with cross-functional teams to integrate geospatial data with machine learning insights, facilitating decision support tools for stakeholders. Engaged in strategic discussions to align project objectives with technological advancements in geospatial analytics.
- Rapidly prototyped web apps based on stakeholder requirements and scaled cloud infrastructure to meet project specifications.