POOJA BARALU UMESH

LinkedIn GitHub Blog

Bay Area, CA, USA +16692491322

poojabumesh@gmail.com

SUMMARY

Machine Learning Engineer with 4+ years of experience and MS in Data Science, including 1+ year building and deploying production-grade ML systems in e-commerce, supply chain, and computer vision domains. Skilled in LLMs, vector embeddings, and end-to-end MLOps (AWS, Spark, Airflow). Proven ability to design, deploy, and optimize scalable AI solutions that reduce latency, improve accuracy, and deliver measurable business impact.

SKILLS

Programming Languages: Python, SQL
Machine Learning: Scikit-learn, TensorFlow, PyTorch, MLflow, Hugging Face Transformers
ML Concepts: Vector Embeddings, LLMs, Prompt Engineering, NLP, Feature Engineering, Evaluation Frameworks, Time Series Forecasting, Clustering, Dimensionality Reduction (PCA), Content-Based Recommendation
Statistics & Analysis: ANOVA, Regression, GLM, Hypothesis Testing, Experimental Design
Data Processing & Workflow: Pandas, NumPy, Apache Spark (SparkSQL, DataFrames), Airflow
Cloud Platforms: AWS, GCP
Databases: MongoDB, Pinecone (Vector DB)
Visualization & Apps: Matplotlib, Seaborn, Streamlit
APIs & Tools: SpaCy, NLTK, Cohere API

PROFESSIONAL EXPERIENCE

Drinks, San Jose, CA Oct 2024 - July 2025
Machine Learning Engineer, Part-time **Blog**

- Designed & deployed a **Retrieval-Augmented Generation (RAG)** product search agent using **vector embeddings, OpenAI LLM**, and **Pinecone vector DB.**
- Built intent classification (recommendation/info/conversational) with **cosine similarity & Levenshtein distance** for accurate query routing.
- Integrated **Hugging Face cross-encoder reranker** to improve semantic match accuracy and implemented product filtration to reduce duplicates.
- Orchestrated **AWS Lambda, Cohere Reranker,** and **S3 storage** for 10k+ embeddings, cutting response time from **~2 min to <30 sec.**
- Created evaluation framework with 250 curated queries for continuous performance monitoring; deployed solution site-wide via API.
- Presented the solution to the Co-founder and CTO; successfully deployed company-wide for end-user adoption and integrated it seamlessly into the e-commerce website through API orchestration.

Harmony Food Pvt. Ltd., India Mar 2022 - Jan 2024
Software Engineer

- Designed and implemented **ETL pipelines** in Python/SQL to ingest production & quality-control data, cutting reporting time by 60%.
- Built **data validation scripts** and automated alerts for anomalies in ingredient supply and production KPIs, improving issue resolution speed.
- Developed **interactive dashboards** (Power BI/Tableau) for leadership to monitor real-time inventory, demand forecasts, and supplier performance.
- Integrated APIs from suppliers and logistics partners into the central database to enhance supply chain visibility.
- Collaborated with QA and operations teams to translate business needs into technical specifications for analytics and reporting systems.

ABInBev, India Oct 2020 - Jan 2022
Assistant Manager

- Led a cross-functional analytics project that reduced extract loss by 9% and water usage by 5% through **time-series anomaly detection** on brewhouse sensor data.
- Developed and maintained **SQL data pipelines** to track Overall Equipment Effectiveness (OEE) metrics, enabling near real-time performance monitoring and driving a 25% productivity increase.
- Automated **report generation** and KPI tracking by integrating data from MES (Manufacturing Execution Systems) into centralized dashboards.
- Collaborated with software and IT teams to scope, test, and deploy new process-monitoring tools on production lines.
- Implemented **quality control improvements**, cutting customer complaints by 7% through root-cause analysis and system alerts.

EDUCATION

University of San Francisco, **SF, CA, USA**
MS in Data Science July 2024 - June 2025

PROJECTS

Developed A(I)YE Chef, an end-to-end AI-powered culinary assistant [Github]

- Fine-tuned YOLOv8 on 24k+ images for 120-class ingredient detection (>95% accuracy).
- Integrated Vertex AI Gemini LLM to generate JSON-structured personalised recipes.
- Exported and deployed the YOLOv8 PyTorch model via FastAPI in a Docker container, leveraging GCP Cloud Run for scalable, serverless inference
- Implemented MLflow for artifact logging and model registration, resolving conflicts with YOLO's auto-logging to centralize metric tracking and optimize compute costs through transfer learning and serverless deployment.

Tweet Popularity Predictor – End-to-end ML pipeline for social media analytics [Github]

- Multi-task pipeline for emotion classification (DistilBERT), hashtag generation (GPT-2), and popularity scoring (linear regression).
- Designed retrainable Python package with CLI & API; integrated with Snowflake for storage & dashboards.
- Optimized inference speed via batch processing, reducing runtime for large datasets; added unit test scaffolding and roadmap for FastAPI microservice deployment.

Fine-Tuning Stable Diffusion 2.1 for Domain-Focused Image Generation [Blog]

- Fine-tuned Stable Diffusion 2.1 on curated ArtBench-10 dataset, enabling stylistically coherent image generation aligned with domain-specific artistic prompts and style requirements..
- Designed robust preprocessing pipeline including deduplication via perceptual hashing, CLIP normalization, and UTF-8 text cleaning to ensure high-quality training data.
- Implemented memory-efficient training using WebDataset with 46 sharded tar files and LoRA-based PEFT on A100/4090 GPUs, optimizing resource utilization and training time.
- Evaluated model performance using CLIP similarity scores and human assessment, achieving improved prompt adherence and visual fidelity in generated artwork

Webflix Browse Time Optimization

- Led a collaborative analysis using a two-stage factorial experiment and simulated user data to optimize recommendation settings (Tile Size, Match Score, Preview Length, Type) and minimize user browsing time.
- Applied statistical methods including ANOVA, partial F-tests, Bonferroni correction, and OLS regression to identify significant factors and interactions, determining an optimal configuration (Preview Type: TT, Length: 75s, Score: 72%).
- Delivered a data-driven strategy predicted to reduce mean browsing time by 20% (estimated ~9.98 min), enhancing user engagement.

Movie Recommendation System Pipeline [Github]

- Developed components of an automated data pipeline using Airflow, MongoDB, GCS, and Spark for content-based movie recommendation systems.
- Ingested data from TMDB APIs, performed data transformations (joins, aggregations) in MongoDB, and enabled scalable analytics using Spark DataFrames and SparkSQL.